

TC

EGE ÜNİVERSİTESİ

Fen Bilimleri Enstitüsü

**ÇOKLU ETİKETLEME İLE METİN BÖLÜTLEME İÇİN
SEZGİ ÜSTÜ BİR YÖNTEM**

Selcan KAYAHAN

DOKTORA TEZİ

**Tez Danışmanı: Prof. Dr. Urfat NURİYEV
İkinci Tez Danışmanı: Doç. Dr. Korhan GÜNEL**

**Matematik Anabilim Dalı
Bilgisayar Bilimleri Bilim Dalı**

Sunuş Tarihi: 20.10.2023

Bornova-İZMİR

2023

Selcan KAYAHAN tarafından doktora tezi olarak sunulan “Çoklu Etiketleme ile Metin Bölütleme İçin Sezgi Üstü Bir Yöntem” başlıklı bu çalışma EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliği ile EÜ Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi'nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve 20/10/2023 tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı bulunmuştur.

Jüri Üyeleri:

İmza

Jüri Başkanı	:
Raportör Üye	:
Üye	:
Üye	:
Üye	:

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

ETİK KURALLARA UYGUNLUK BEYANI

EÜ Lisansüstü Eğitim ve Öğretim Yönetmeliğinin ilgili hükümleri uyarınca Doktora Tezi olarak sunduğum “Çoklu Etiketleme ile Metin Bölütleme İçin Sezgi Üstü Bir Yöntem” başlıklı bu tezin kendi çalışmam olduğunu, sunduğum tüm sonuç, doküman, bilgi ve belgeleri bizzat ve bu tez çalışması kapsamında elde ettiğimi, bu tez çalışmasıyla elde edilmeyen bütün bilgi ve yorumlara atıf yaptığımı ve bunları kaynaklar listesinde usulüne uygun olarak verdiğimi, tez çalışması ve yazımı sırasında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, bu tezin herhangi bir bölümünü bu üniversite veya diğer bir üniversitede başka bir tez çalışması içinde sunmadığımı, bu tezin planlanmasından yazımına kadar bütün safhalarda bilimsel etik kurallarına uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul edeceğimi beyan ederim.

20 / 10 / 2023

İmzası

Selcan KAYAHAN

ÖZET**ÇOKLU ETİKETLEME İLE METİN BÖLÜTLEME İÇİN****SEZGİ ÜSTÜ BİR YÖNTEM**

KAYAHAN, Selcan

Doktora Tezi, Matematik Anabilim Dalı

Tez Danışmanı: Prof. Dr. Urfat NURİYEV

İkinci Tez Danışmanı: Doç. Dr. Korhan GÜNEL

Ekim 2023, 60 sayfa

Bu tezde, öncelikle eğitim içeriklerinden otomatik olarak öğretim kavramları tespit edilmiştir. Ardından metinsel veri formundaki eğitim içeriği, anlamsal bütünlüğü olan ve birbiriyle örtüşen metin bloklarına bölünmüş ve bu bloklardaki paragrafların birden fazla öğretim kavramıyla etiketlenmesi hedeflenmiştir. Tez, Millî Eğitim Bakanlığı tarafından okutulan Sanat Tarihi, Türk Kültür ve Medeniyet Tarihi kitaplarını temel almaktadır. Kitaplarda yer alan metinler üzerinde sezgisel ve sezgi üstü kümeleme yöntemleri, doğal dil işleme yöntemleri uygulanmış ve paragrafların hangi öğretim kavramıyla ilişkilendirildiği belirlenmeye çalışılmıştır. Bu amaç doğrultusunda, metin bloklarından öznitelik vektörleri çıkarılmış ve bu vektörler üzerine Temel Bileşen Analizi (PCA) uygulanmıştır. Daha sonra Parçacık Sürü Optimizasyonu (PSO) yöntemi kullanılarak kümeleme yapılmıştır. Ayrıca, önerilen modelin başarımını değerlendirmek için kitaplar içinde sunulan öğretim kavramları ile uzmanlar tarafından eşleştirilmiş bölütlendirilmiş metin blokları kullanılmıştır. Ağırlıklandırılmış ortalama karesel hata değerleri, sistemin çıktıları ile uzman görüşleri karşılaştırılarak hesaplanmıştır. Sonuçlar incelendiğinde, eğitim içeriklerinin çoklu olarak öğretim kavramları ile etiketlenmiş bir şekilde metin bloklarına bölünebileceği umudu ortaya çıkmıştır.

Anahtar sözcükler: Çoklu etiketleme, parçacık sürü optimizasyonu, sezgisel metin bölütleme, eğitim teknolojileri



ABSTRACT**A META HEURISTIC METHOD FOR TEXT SEGMENTATION****WITH MULTI-LABELING**

KAYAHAN, Selcan

PhD in Mathematics

Supervisor: Prof. Dr. Urfat NURİYEV

Co-Supervisor: Assoc. Prof. Dr. Korhan GÜNEL

October 2023, 60 pages

In this thesis, first, the learning concepts have been extracted from an educational content automatically. Next, the educational content in textual format was divided into text blocks that have semantic integrity and overlap with each other, and the paragraphs in these blocks were primarily aimed to be labeled with more than one learning concept. The thesis is based on the History of Art and History of Turkish Culture and Civilization books taught in schools affiliated with the Ministry of National Education. Heuristic and meta heuristic clustering methods and natural language processing were applied on the books and it was tried to determine which learning concept each paragraph was associated with. For this purpose, feature vectors were extracted from text blocks and Principal Component Analysis (PCA) was applied on these vectors. Then, clustering was carried out using the Particle Swarm Optimization (PSO) method. Additionally, segmented text blocks paired with the learning concepts presented in the books by experts were used to evaluate the performance of the proposed model. Weighted mean square error values were calculated by comparing the outputs of the system with expert opinions. When the results were examined, the hope emerged that educational content could be divided into text blocks multi labeled with learning concepts.

Keywords: Multi-labeling, heuristic text segmentation, particle swarm optimisation, educational technology

ÖNSÖZ

Doğal Dil İşleme, yapay zekanın önemli alt dallarından biridir ve uzun zamandır arařtırmacıların ilgisini çekmektedir. Metin işleme yöntemlerinin eğitim teknolojileri alanında kullanılması fikri, doktora ders aşamam sırasında ortaya çıktı. Başta danışman hocalarım olmak üzere ders hocalarımın da desteęi ile bu fikrin doktora tezine dönüşmesi, bir öğretmen olarak beni mutlu ediyor. Özel ilgi alanım olan sanatın tezimde yer alması da benim için ayrı bir mutluluk. Arařtırmamın ve çalışmalarımın ilgili literatürlere katkı sağlamasını dilerim.

İZMİR

20/10/2023

Selcan KAYAHAN



İÇİNDEKİLER

	<u>Sayfa</u>
İÇ KAPAK	ii
KABUL ONAY SAYFASI	iv
ETİK KURALLARA UYGUNLUK BEYANI.....	vi
ÖZET	viii
ABSTRACT	x
ÖNSÖZ.....	xii
İÇİNDEKİLER DİZİNİ.....	xiv
ŞEKİLLER DİZİNİ	xvi
TABLolar DİZİNİ.....	xviii
SİMGELER VE KISALTMALAR DİZİNİ.....	xx
1. GİRİŞ.....	1
1.1 Konu ve Amaç	1
1.2 Literatür Özeti.....	2
2. MATERYAL VE METOT	10
2.1 Önişlem.....	11
2.2 En Uzun Gövde Tespiti	11
2.3 n-gram Çıkarımı	13

İÇİNDEKİLER (devam)

2.4 n-gram Filtreleme.....	14
2.5 Öznitelik Vektörlerinin Oluşturulması.....	15
2.6 Normalizasyon	15
2.7 Boyut İndirgeme	16
2.8 Kümeleme	17
2.8.1 K-Means.....	17
2.8.2 Parçacık Sürü Optimizasyonu.....	18
3. DENEYSEL ÇALIŞMALAR.....	23
4. SONUÇ VE TARTIŞMA	28
KAYNAKLAR DİZİNİ	30
TEŞEKKÜR.....	36
ÖZGEÇMİŞ	37

ŞEKİLLER DİZİNİ

<u>Şekil</u>	<u>Sayfa</u>
2.1 Eğitim dokümanlarının öğretim kavramları ile çoklu etiketleme yaparak etiketlenmesi.....	8
2.2 Ortak paragraf içeren metin bloklarının ayrıştırılması	10
2.3 PSO (Parçacık Sürü Optimizasyonu) algoritmasının akış şeması.....	19
2.4 Parçacıkların her bir iterasyonda yeni konuma yönelimi.....	20
3.1 PCA (Principal Component Analysis) uygulandıktan sonra metin bloklarının düzlemde konumlandırılması	23
3.2 Ortak paragraf içeren metin bloklarının etiketlenmesi	23



TABLolar DİZİNİ

<u>Tablo</u>	<u>Sayfa</u>
3.1 En uzun gövde tespiti aşamasında uygulanan işlemler ve açıklamaları ...	22
3.2 Doküman hakkında bazı istatistiki bilgiler.....	23





SİMGELER VE KISALTMALAR DİZİNİ

<u>Simgeler</u>	<u>Açıklama</u>
n -gram	n 'li dizilim
n	Aday öğretim kavramındaki sözcük sayısı
t_j	Aday öğretim kavramı
$f(t_j)$	t_j 'nin (aday öğretim kavramı) eğitim içeriğinde geçme sayısı
$tf(t_j)$	n -li sözcük diziliminden oluşan t_j aday öğretim kavramının görüntülenme sıklığının eğitim içeriğindeki tüm n -li sözcük dizilimlerinin toplam frekansına oranı
$df(t_j)$	t_j aday öğretim kavramını içeren metin bloklarının sayısı
$itbf(t_j)$	Ters metin bloğu frekansı
$tbf - itbf(t_j)$	Metin bloğu frekansı - Ters metin bloğu frekansı
x'	Normalizasyon değeri
M_k	k kümesi merkezi
E_k^2	Küme içi değişimler için toplam karesel hata
P	Parçacık sürü optimizasyonunda olası çözüm kümesi

Kısaltmalar

K-NN	K-En Yakın Komşu (K-Nearest Neighbour)
PCA	Temel Bileşenler Analizi (Principal Component Analysis)
PSO	Parçacık Sürü Optimizasyonu (Particle Swarm Optimization)



1. GİRİŞ

Bu tezde, eğitim içerikleri üzerinde sezgisel ve sezgi üstü metin bölütleme yaklaşımları uygulanmıştır. Tez 4 bölüm, ek ve kaynaklar listesinden oluşmaktadır.

Birinci bölümde (**Giriş**) ilk olarak tezin amacı ve kapsamı açıklanmış, ardından konu ile ilgili güncel bir literatür taraması sunulmuştur.

İkinci bölümde (**Materyal ve Metot**) tezde kullanılan yöntemler anlatılmıştır.

Üçüncü bölümde (**Deneysel Çalışmalar**) yapılan uygulamalar açıklanmıştır.

Son olarak dördüncü bölümde (**Sonuç ve Tartışma**) ise tez çalışmasının sonuçları ortaya koyulmuş ve tez çalışması ile ilgili öneriler sunulmuştur.

1.1 KONU VE AMAÇ

Bilgi, nesnel gerçekliğin insan bilincinde yansımaları, oluşumu ve bu oluşumun bir sonucudur (Lucaks, 1951). Kavram ise; nesnelere veya olayların ortak özelliklerini tek bir ortak ad altında toplayan genel bir tanımlamadır (Türk Dil Kurumu, 2021). İnsanlar yaşamlarında, bilgileri kendi temel zihinsel oluşumları içinde var etmek için kavramlara gereksinim duyarlar. Nesnelere, olgular, olaylar insan zihninde kavramlar yolu ile anlaşılır ve ayırt edilir. Bozkurt (2018), her bir deneyimin, bireyin kavramlarını, sahnelerini, şemalarını ve anlam ağlarını yeniden yapılandırmak anlamına geldiğini vurgulamıştır. Okullarda kullanılan ders kitapları gözden geçirildiğinde, her ünite veya kitap için bir 'kavramlar listesi' sunulduğu görülür. Bu liste, öğrencilerin bir bölüme ya da üniteye başlamadan önce o konunun hangi kavramlarla ilgili olduğunu anlamalarına yardımcı olur. Bu sayede öğrenciler, çalışacakları konunun genel hatlarıyla ne ile ilgili olduğunu bilerek çalışmaya başlarlar. Neleri öğreneceğinin bilincinde olan öğrenen, zihnindeki sahne, şema ve anlam ağlarını yapılandırmaya başlar ve konunun öğrenilmesi kolaylaşır. Görüldüğü üzere ders kitaplarında bulunan kavramların doğru, tam ve konuyu genel olarak kapsayacak şekilde verilmesi, öğrenmenin kolaylıkla gerçekleşebilmesi bakımından önem arz etmektedir.

Bu tezde, sezgisel segmentasyon yaklaşımı ile metin içerikli bir eğitim belgesini, bu dokümandan otomatik olarak tespit edilen birden fazla öğretim kavramı ile etiketlenmiş metin bloklarına ayırıştıran bir model önerilmektedir. Bu çerçevede, önce, sezgisel ve sezgi-üstü optimizasyon teknikleri, makine öğrenimi ve istatistiksel dil modelleri ile bahsi geçen eğitim içeriğinden temel öğretim kavramlarının bir alt kümesi çıkarılmıştır. Sonrasında belge, çakışan metin bloklarına ayrılmış ve her paragrafın tespit edilen öğretim kavramları ile etiketlenmesi bu metin blokları üzerinden sağlanmıştır.

Daha genel bir perspektiften bakıldığında ise, önerilen model, öğrenenlere özelleştirilmiş bir arama deneyimi sunan bir eğitim dokümanı arama motoru modeli olarak düşünülebilir. Bu model, öğrenenin belirli bir kavramı veya konuyu aradığında, ilgili metin bloklarını dikkate alarak daha kapsamlı ve anlamlı sonuçlar sağlayabilir. Örneğin, bir öğrenci "koşul ifadeleri" hakkında bilgi ararken, bu model, "koşullar" kavramını içeren dokümanlara yönlendirmek yerine, koşul ifadeleri ile ilgili tüm metin bloklarını sunarak daha zengin bir bilgi kaynağı sunabilir. Bu yaklaşım, öğrencilerin daha spesifik bilgilere ulaşmalarını kolaylaştırabilir ve öğrenme deneyimlerini daha etkili hale getirebilir.

1.2 LİTERATÜR ÖZETİ

Dünya genelinde hızla artan dijitalleşme, eğitim alanında da büyük veri miktarlarının artmasına neden olmaktadır. Öğrenciler, öğretmenler ve eğitim ortamlarından elde edilen bu verilerin analizi, eğitim bilimleri ve eğitim teknolojileri alanında kritik bir adımdır. Bu analizlerin sonuçlarına dayalı olarak öğrencilere öğrenme süreçlerini daha verimli bir şekilde yönlendirebilmek için rehberlik sağlanabilir. Luckin ve Çukurova (2019), bireyselleştirilmiş eğitim sistemleri için alt yapı oluşturulmasına ve büyük verinin analiz edilmesine yardımcı olabilecek yapay zeka algoritmalarının önemini vurgulamaktadır. Ayrıca, Renz ve Hilbig (2020), yapay zeka içeren eğitim teknolojilerinin gereksinimlerini tanımlamışlardır. Bu çalışmalar, eğitim alanındaki veri analizi ve yapay zeka kullanımının önemini vurgulayarak öğrenme süreçlerini daha etkili hale getirmeye yönelik fırsatları öne çıkarmaktadır.

Ouyang ve Jiao (2021), eğitim teknolojilerinde yapay zeka kullanımını değerlendiren üç farklı paradigma önermişlerdir. İlk yaklaşımlarında, yapay zekanın öğrenme süreçlerinde sadece bir araç olarak kullanılmasını ve öğrenenlerin bu araçları kullanıcı olarak görmesini

savunmuşlardır. İkinci yaklaşımlarında ise, öğrenenlerin yapay zeka ile birlikte etkileşimde bulunarak, sistem içinde daha aktif bir rol oynamalarını önermişlerdir. Bu bakış açısıyla yapay zekanın yaklaşımlarının, eğitim teknolojilerine entegre edilmesi eğitim süreçlerinde destek işlevi görür. Son olarak, öğrenen davranışlarıyla (öğrenenlerden elde edilen girdilerle) eğitilen yapay zeka sistemleri herkesin kendi öğrenme seviyelerini ve öğrenme stillerinin geliştirmek için kullandığı bir araç şeklinde sunulduğu bir sistemden söz etmişlerdir. Bu üç yaklaşım, yapay zekanın eğitimde nasıl kullanılabileceği konusunda farklı perspektifler sunmaktadır. Zhai ve arkadaşları (2021), eğitim alanında kullanılan yapay zeka yöntemlerini içeren bir literatür incelemesi sunarak, bu alandaki gelişmeleri özetlemişlerdir. Nabiyev ve Erümit (2020), eğitim ortamlarında yapay zeka kullanımının teknik ve pedagojik sonuçlarını ele almış ve bu konudaki literatürün eksikliklerine dikkat çekmişlerdir. Akdeniz ve Özdiç (2021), Türkiye'de uygulanan makine öğrenimi tekniklerini ve eğitim teknolojileriyle ilgili çalışmaları inceleyen sistematik bir literatür çalışması gerçekleştirmişlerdir. Bu çalışma, Türkiye'deki eğitim teknolojileri alanındaki araştırmaların mevcut durumunu sunmaktadır.

ChatGPT, OpenAI tarafından geliştirilmiş olan GPT-4 mimarisine dayanan ve geniş bir metin veri kümesi üzerinde eğitilmiş bir doğal dil işleme modelidir. Bu sayede çoklu dil desteği, anlam ve dilbilgisi kurallarına uygunluk gibi özellikleri destekler. Doğal dil işleme teknikleri arasında kelime dağılımı, belirli kelime gruplarına odaklanma, kelime köklerinin belirlenmesi, cümle yapısı analizi ve sözcük anlamlarının karşılaştırılması yer alır. Yürütülen çalışma ile ilgili ChatGPT'ye “Türkçe bir metinde paragrafları birden fazla kavramla etiketleyebilir misin?” sorusu sorulmuş ve şu cevap alınmıştır: “Evet, Türkçe bir metinde paragrafları birden fazla kavramla etiketlemek mümkündür. Örneğin, bir paragraf hem tarih hem de kültür kavramlarını içerebilir veya hem ekonomi hem de siyaset kavramlarını içerebilir. Burada önemli olan, paragrafların içeriğine göre uygun kavramlar seçmektir.” Örnek olarak, aşağıdaki paragraf ele alınabilir:

“Osmanlı İmparatorluğu'nun son dönemlerinde, ekonomik kriz ve siyasi istikrarsızlık ülkeyi etkisi altına almıştı. Halkın büyük çoğunluğu yoksulluk içinde yaşarken, azınlık gruplar ise ayrıcalıklı konumlarını koruyorlardı.” Bu paragraf hem ekonomi hem de siyaset kavramlarıyla etiketlenebilir. Ayrıca, kültür kavramı da paragrafta yer almaktadır, çünkü Osmanlı İmparatorluğu'nun son dönemi Türk kültürü için önemli bir dönemdir.

Tezde önerilen model, çoklu etiketlemeli metin bölütleme yaklaşımının eğitim teknolojilerinde kullanılması, makine öğrenimi alanında doküman sınıflandırma ve etiketleme

üzerine yapılan arařtırmalara yeni bir örnek sunmaktadır. Doküman sınıflandırma ve etiketleme konusu literatürde yaygın olarak ele alınmış bir konu olmasına rağmen, çoklu etiketleme, klasik sınıflandırma yaklaşımlarından farklıdır. Burkhardt ve Kramer (2018) tarafından belirtildiđi gibi, çoklu etiketleme problemleri, problem dönüşüm yöntemleri ve uyarlanabilir algoritmalar olacak şekilde iki ana kategori altında incelenebilir. Problem dönüşüm yaklaşımları, veriyi birden fazla sınıflandırıcı ile etiketlemenin mantığına dayanır. Bu yaklaşım, veriyi çeşitli şekillerde işleyerek her bir etiketi tahmin etmek için farklı sınıflandırıcıları kullanmayı içerebilir. Uyarlanabilir algoritma yaklaşımlarında ise, tek bir sınıflandırıcı kullanılır ve bu sınıflandırıcı, algoritmanın veriye göre çalışması için gerekli olan parametre değerlerini güncelleyerek ilerler. Bu yaklaşım, modelin öğrenme yeteneđini artırmak ve daha iyi sonuçlar elde etmek için kullanılır. Bu tezde önerilen model, çoklu etiketlemeli metin bölütleme yaklaşımını eğitim teknolojilerine uyarlamış ve bu alandaki literatüre yeni bir katkı sağlamıştır. Bu yaklaşımın, eğitim materyallerinin daha etkili bir şekilde anlaşılması ve özelleştirilmesi gibi uygulamalarda kullanılabilirliđi bulunmaktadır.

Ferrandin ve Cerri (2023), çoklu etiketlemeli yaklaşımların son yıllarda biyoenformatik, sağlık, müzik sınıflandırma, metin sınıflandırma, dolandırıcılık tespiti gibi birçok farklı gerçek dünya probleminin çözümü için kullanıldığını belirtmişlerdir. Yine Jia ve Zhang (2023), gerçek dünyada zengin anlam içeriđine sahip nesnelerin bulunduđunu ve bu nesnelerin özelliklerinin klasik bir yaklaşımla tek bir etiketle etiketlenmesi yerine daha karmaşık bir şekilde karakterize edilmesi gerektiđini belirterek çoklu etiketlemenin önemine vurgu yapmışlardır. Tarekegn ve arkadaşları (2021) çalışmalarında, çoklu etiketleme yaklaşımlarının dengeli olmayan veri setlerinde uygulandıđı bir literatür çalışması sunmuşlardır. Benzer olarak Moyano ve arkadaşları da (2018), çoklu etiketlemeli sınıflandırıcıları içeren bir literatür taraması hazırlamışlardır. Bu iki çalışmada vurgulanan ana konu, veriyi tek etiketle işaretleme ile birden çok etiketle sınıflandırma yaklaşımları arasındaki farklılıklar mevcuttur ve çalışmada çoklu etiketlemede ortaya çıkan zorluklar vurgulanmıştır. Tarekegn ve arkadaşları (2021) çoklu etiketlendirme metotlarını; yeniden örnekleme (resampling), adapte edilebilir sınıflandırma (adaptable classification), topluluk ve maliyet duyarlı (ensemble and cost-sensitive) yaklaşımlar olacak şekilde dört farklı başlıkta gruplandırmışlardır (Kayahan ve arkadaşları, 2022). Literatürde yer alan yaklaşımların performans analizini yapmak üzere sunulan ölçütler ise örnek (example), etiket (label) ve sıralama (ranking) temelli ölçütler olacak şekilde üç ana kategoride sıralamışlardır. Bu ölçütler, çoklu etiketleme sınıflandırıcılarının performansının objektif bir şekilde

değerlendirilmesine yardımcı olur. Tarekegn ve arkadaşları (2021), dengeli olmayan veri kümelerinde çoklu etiketleme yaklaşımları ile ilgili bir inceleme makalesi sunarak, özellikle tekli etiketleme ile çoklu etiketleme arasındaki farkları ve bu alandaki zorlukları vurgulamışlardır. Aynı şekilde, Moyano ve arkadaşları (2018), çoklu etiketleme sınıflandırıcıları için bir literatür tarama makalesi sunmuş ve aynı konuya odaklanmışlardır. Bu iki çalışmada temel vurgu, veriyi tek etiketleme ile çoklu etiketleme yaklaşımları arasındaki farklar ve bu farklılıkların neden olduğu zorluklardır.

Kumar ve arkadaşları (2018), hiyerarşik yerleştirme mantığı ile çoklu etiketlemeli bir sınıflandırma tekniği olan MLC-HMF'yi ortaya koymuşlar ve geliştirdikleri yöntemi çeşitli uygulama alanlarında test verisi olarak sunulan veri setleri üzerinde test etmişlerdir. Çalışmalarında dikkat çeken önemli bir nokta, özellikle eğitim ile ilgili veri setleri için elde edilen doğruluk (accuracy) değeri oranının $29,4 \pm 2,1$ ile sınırlı olmasıdır.

Liu ve arkadaşları (2005), blog yazıları için uyguladıkları duygu analizi ile ilgili çalışmasında çoklu etiketleme yaklaşımını kullanmış ve %34,4'lük bir doğruluk (accuracy) oranı elde etmişlerdir. Bu, metinlerin duygusal içeriğini çoklu etiketlerle sınıflandırmak için kullanılan yaklaşımın performansını göstermektedir. Tarekegn ve arkadaşlarının (2021) literatür inceleme çalışmasında ise çoklu etiketleme sınıflandırıcılarının performansını değerlendirmek için farklı ölçütler kullanılmıştır. Bu ölçütler şu alt başlıkları içermektedir.

Hamming kaybı (Hamming loss): Etiketlerin tahmin edilen ve gerçek etiketler arasındaki farkları ölçer ve düşük bir değer aranır.

Alt küme doğruluk (subset accuracy): Tüm etiketlerin doğru bir şekilde tahmin edilmesinin yüzdesini hesaplar.

Örnek tabanlı F1 (example-based F1) skoru: Precision (kesinlik) ile recall (hassasiyet) skorlarının geometrik ortalamasıyla performansı değerlendirir.

Mikro ve makro F1 skorları: F1 skorları, sınıflandırma performansını ölçen bir metrik olarak kullanılır.

Ortalama kesinlik (average precision): Etiketlerin tahmin edilme kesinliğini ölçer.

Kapsama (coverage): Etiketlerin tahmin edilme kapsamını ölçer.

Bir hata (one error): Her etiket için doğru tahminin yapılmasının zorunlu olduğu bir ölçüttür.

Bu ölçütler, çoklu etiketleme sınıflandırıcılarının performansını farklı yönlerden değerlendirmek için kullanılır ve hangi metriğin daha uygun olduğu, veri seti ve uygulama bağlamına bağlı olarak değişebilir.

Lee ve arkadaşları (2019), doküman sınıflandırma için öznitelik seçimi amacıyla memetik bir arama algoritması geliştirmişler ve bu yaklaşımı kullanarak dokümanları farklı genel kategorilere (örneğin Sanat, Bilgisayar, Eğitim) sınıflandırmışlardır. Önerdikleri yaklaşımı, farklı türdeki (filtre ve sarmal tabanlı) öznitelik seçim yöntemleri ile karşılaştırmışlardır. Her bir kategori için ayrı ayrı doğruluk değerleri hesaplanarak yöntemin performansı değerlendirilmiştir. Elde ettikleri doğruluk oranlarının %13,22 ile %53,45 arasında değiştiği görülmektedir. Kimm, Lee ve Seok (2022), ekonomi haber makalelerinden ekonomik kavramların etiketlerini çıkarmak için çok etiketli bir metin sınıflandırma yaklaşımı uygulamışlardır ve bu yaklaşımla elde ettikleri doğruluk oranını %58 olarak bildirmişlerdir. Her iki çalışma da doküman sınıflandırma ve etiketleme alanında özgün yaklaşımlar geliştirmiş ve farklı metrikler kullanarak performanslarını değerlendirmişlerdir. Elde edilen sonuçlar, bu yaklaşımların farklı veri kümesi ve uygulama bağlamına göre performanslarının değişebileceğini göstermektedir.

Yang ve Liu (2019), çoklu etiketli metin sınıflandırma için seri kod çözme ve paralel kodlama yaklaşımını temel alan bir yöntem ortaya koymuşlardır. Bu yöntem, evrimsel yapay sinir ağlarını kullanarak, verilen dokümandan global etkileşim ve yerel komşuluk bilgisini birlikte çıkararak bir kodlayıcıyı birleştirir. Çalışma, üç farklı veri seti üzerinde denenmiş ve her veri seti için mikro-F1 değerleri sırasıyla 0,893, 0,725 ve 0,825 olarak bulunmuştur. Bu sonuçlar, çoklu etiketlemeli doküman sınıflandırıcılarının performansının kabul edilebilir düzeylerde olduğunu göstermektedir. Yang ve Liu'nun geliştirdiği yöntem, metin verileri üzerinde çoklu etiketleme yapmak için etkili bir yaklaşım sunmaktadır.

Aljedani ve arkadaşları (2020), hiyerarşik bir şekilde çoklu etiketleme yöntemi olan HOMER (Hierarchy Of Multilabel ClassifiER) algoritmasını Arapça metinler üzerinde etiketleme çalışması için uygulamışlardır. Bu çalışma, birden fazla sınıflandırıcının kullanıldığı HOMER algoritmasının parametrelerini optimize etme amacını taşımaktadır. Çalışma sonuçlarına göre, Aljedani ve ekibi, çoklu etiketleme için %75.8 doğruluk oranı ve ortalama olarak %85.3 mikro-F1 skoru elde etmişlerdir. Bu sonuçlar, Arapça metinler

üzerinde etiketleme görevlerinde HOMER algoritmasının etkili bir şekilde kullanılabilceğini göstermektedir. Bu tür çalışmalar, çoklu etiketleme yöntemlerinin farklı dillerdeki metin verileri üzerinde uygulanabilirliğini değerlendirmek için önemlidir ve metin madenciliği alanında dil bağımsız etiketleme yaklaşımlarının geliştirilmesine katkı sağlayabilir.

Çoklu etiketleme yaklaşımları, geleneksel tek etiketli sınıflandırma ve kümeleme yöntemlerine kıyasla bazı zorluklarla karşılaşır. Bu zorluklar, etiket kombinasyonlarının çoklu ve farklı olabileceği, veri setlerinin dengesizliği, eğitim verilerinin eksik veya hatalı etiketlenmiş olabileceği gibi faktörlerden kaynaklanabilir. Bu nedenle, çoklu etiketleme yöntemlerinin başarı oranları, veri seti, problem alanı ve kullanılan yaklaşıma bağlı olarak değişkenlik gösterebilir. Ayrıca, eğitim teknolojileri alanında çoklu etiketleme yöntemlerinin uygulandığı çalışmaların sınırlı olması, bu alandaki potansiyel uygulamaların henüz tam anlamıyla keşfedilmediğini gösterebilir. Eğitim teknolojileri, öğrencilerin öğrenme süreçlerini iyileştirmek ve kişiselleştirmek için veri analitiği ve yapay zeka tekniklerini kullanma potansiyeline sahiptir. Bu bağlamda çoklu etiketleme yaklaşımları, öğrencilerin ilgi alanlarını, güçlü ve zayıf yönlerini daha iyi anlamak ve onlara özelleştirilmiş öğrenme deneyimleri sunmak için kullanılabilir. Gelecekte, eğitim teknolojileri alanında daha fazla çoklu etiketleme çalışması ve uygulamasının geliştirileceği öngörülebilir. Bu alandaki araştırmaların ve uygulamaların geliştirilmesi, öğrencilerin eğitim süreçlerini daha etkili hale getirme potansiyeline sahiptir.

Ayrıca, doküman sınıflandırma uygulamalarının neredeyse tamamı, bir derlem altında birleştirilmiş doküman koleksiyonları üzerinde gerçekleştirilirken, tek doküman üzerinde yapılan doğal dil işleme uygulamaları oldukça nadirdir. Bu tez, geleneksel doküman sınıflandırma yöntemlerini aşarak, tek bir dokümanın içeriğini analiz ederek otomatik öğretim kavramları ile doküman içindeki metin bloklarını etiketleme konusunda özgün bir adım atmıştır. Tez çalışması sırasında, metin bloklarının birden fazla kavramla etiketlenmesi fikri öne çıkmıştır. Ancak bu yaklaşımın bir zorluğu vardır; dokümandaki metin bloklarının içerik analizini yapmak, dokümanın tamamında içerik analizini yapmaktan daha zor bir görevdir. Bunun nedeni, metin bloklarında yer alan sözcük sayısının çok daha sınırlı olmasıdır. Bu problemin üstesinden gelebilmek adına, tezde dokümanların birbiriyle iç içe geçmiş paragraflar içeren alt metin bloklarına ayrılması önerilmektedir. Bu tür bir yaklaşımın literatürde henüz örneklere pek rastlanmadığı görülmektedir. Bu açıdan bakıldığında önerilen fikrin, doğal dil işleme, makine öğrenmesi ve eğitim teknolojilerinde yapay zeka uygulamaları alanlarında literatüre özgün bir katkı sunabileceği düşünülmektedir.

Doğal Dil İşleme, yazılı metinlerin analizi ve işlenmesine odaklanan yapay zekanın bir alt dalıdır. Bu alanın ele aldığı konulardan biri de metin bölütlemedir. Metin bölütleme, bir doküman içinde birbirleriyle bağlantılı ve bir bütünlük içeren metin bloklarını tanımlama işidir (Deepak ve arkadaşları, 2012). Daha anlaşılır ifadeyle, metin bölütleme, özellikle belirli konulara ait ardışık bölümlerin sınırlarını belirleyerek, büyük bir metin içinde yer alan bölümlerin anlamlandırılması ve anlambilimsel olarak bölümler arasındaki farklılıkların ortaya konmasını sağlar (Pak ve Teh, 2017). Beeferman ve arkadaşları (1999), metinsel veriler için segmentasyon problemini karmaşık bir sorun olarak tanımlamış ve metni tutarlı alt bölümlere bölebilmek için n-gram'ların sıklık değerlerini temel alan istatistiksel modelleri incelemişlerdir (Kayahan ve arkadaşları, 2023a).

Metin analizi için belgeyi bölütlemek oldukça yararlı bir yaklaşımdır. Bu yaklaşımın temel nedenlerinden biri, bölünen kısımların belgelerin tamamından daha küçük ve daha tutarlı olmasıdır (Oh ve arkadaşları, 2007). Ayrıca, her bir bölümün analiz ve erişim birimleri olarak kullanılabilir olması da büyük bir avantaj sağlar. Metin bölütlemeyi anlamsal analiz bağlamında kullanan çalışmalar da vardır. Örneğin, arama sonuçlarının iyileştirilmesi için anlambilim kullanılan bir yöntem öneren Hoon, Keong ve Kong (2006) bu konuda öncüdür. Hoon ve Wei (2017) ise arama algoritmalarının geliştirmesi için bilgiyi, değerler olarak adlandırılan segmentler halinde düzenledikleri bir yöntem kullanmışlardır. Ayrıca, görüş madenciliği alanında kullanıcı görüşlerinin hangi yönde eğilim gösterdiğini belirlemek amacıyla Duan ve arkadaşları (2012) metin bölütleme yöntemini uygulamışlardır.

Metin bölütleme, doküman içeriğine otomatik şekilde açıklamalı notlar eklemenin ve metin özetleme çalışmalarının temelini oluşturan önemli bir bileşendir. Doğal dil işleme alanında, dokümanlara açıklayıcı notlar eklemek genellikle metin içinde vurgulanan önemli bölümlerin üst-veri ile etiketlenmesi veya renklendirilmesi yoluyla gerçekleştirilir (Kayahan ve arkadaşları, 2022).

Eğitim materyalleri üzerindeki anlamsal benzerlik ve farklılıkların tespiti, metin bölütleme yöntemleri ile başarılabilir. Metin bölütleme, eğitim dokümanlarını daha küçük anlamlı alt bölümlere bölerek bu alt metin parçaları arasındaki ilişkileri belirleme olanağı sunar. Bu bağlamda, Wentao ve Scheepers (2017), metin böllemenin Çince şiirler üzerinde ritmik okuma hızına ve bilişsel öğrenme sürecine etkisini değerlendirmek için bir çalışma yürütmüşlerdir.

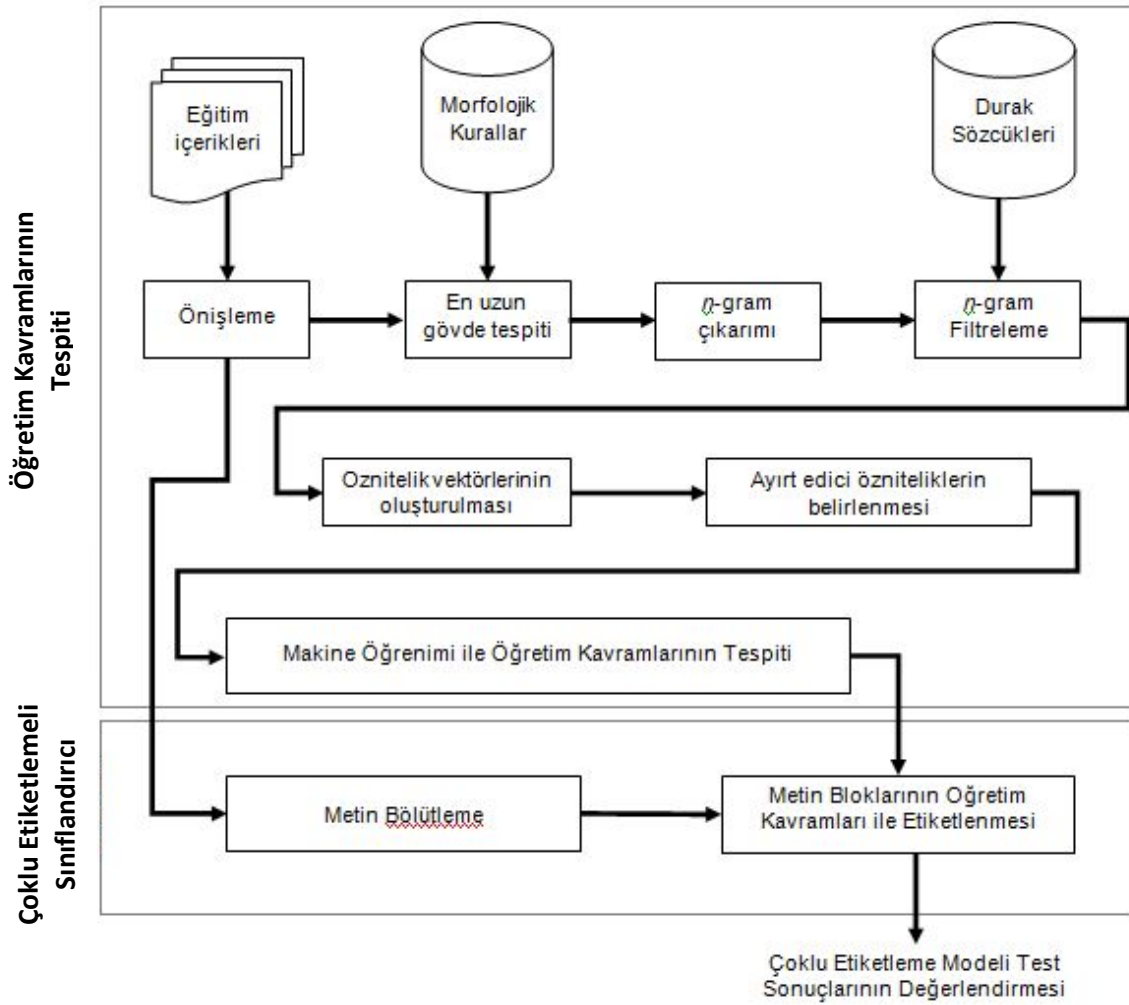
Eđitim dokümanlarını birbirine geen paragraflar Őeklinde metin bloklarına ayırarak her bölümün ayrı ayrı etiketlendirilmesi temeline dayanan bu tez, alt metin bloklarını oluŐturan her bir paragrafı bir ya da daha fazla yine dokümandan otomatik olarak ıkartılan öđretim kavramıyla etiketlemeyi amalamaktadır.

Özellikle üniversite öncesi ders kitapları ve kaynak materyalleri, öđretim kavramlarını ieren yaygın bir Őekilde kullanılır. Bu kavramları belirli bir öđrenme alanına entegre etmek, birçok zaman hem uzmanlar hem de uzman olmayanlar iin zorlu, tartıŐmalı, zaman alıcı ve bazen gereksiz bir süreç olarak görölür (Günel, 2016). Dokümanların veri iŐleme yöntemleri ile analiz edilmesi, eđitim materyallerinin anlamsal analizini yapmada, dokümanların ieriđini anlamada ve öđretim kavramları ile kaynaklar arasındaki iliŐkilerin belirlenmesinde yardımcı olur (Kayahan ve arkadaşları, 2022). Metin bölütleme, dokümanları alt bölümlere ayırarak bu alt bölümler arasındaki iliŐkilerin görölmesine olanak tanır ve eđitim materyalleri üzerindeki anlamsal benzerlikler ile farklılıkların belirlenmesine yardımcı olabilir.

Bu tez, Millî Eđitim Bakanlığı'nın hazırladıđı ve kamuya açık kaynak olarak sunulan kitapların detaylı bir analizini iermektedir. Temel ama, öđrencilerin öđrenme alanlarından bađımsız olarak eđitim dokümanlarındaki metin bloklarını öđretim kavramlarıyla etiketlemektir. Bu dođrultuda, dođal dil iŐleme, makine öđrenimi ve sezgisel kümeleme yaklaŐımları ile ieriklerin kapsamlı bir deđerlendirmesi sunulmaktadır.

2. MATERYAL VE METOT

Bu bölümde ilk olarak eğitim dokümanlarının bölütlenerek öğretim kavramlarıyla etiketlenmesi için geliştirilen sistemin basamakları açıklanacaktır. Sonrasında, K-Means kümeleme yöntemi ve Parçacık Sürü Optimizasyonu (Particle Swarm Optimization, PSO) yaklaşımları açıklanacaktır. Eğitim dokümanlarının parçalanarak öğretim kavramlarıyla etiketlenmesi için geliştirilen modelin adımları Şekil 2.1'de gösterilmiştir.



Şekil 2.1 Eğitim dokümanlarının öğretim kavramları çoklu etiketleme ile etiketlenmesi.

2.1 Önişlem

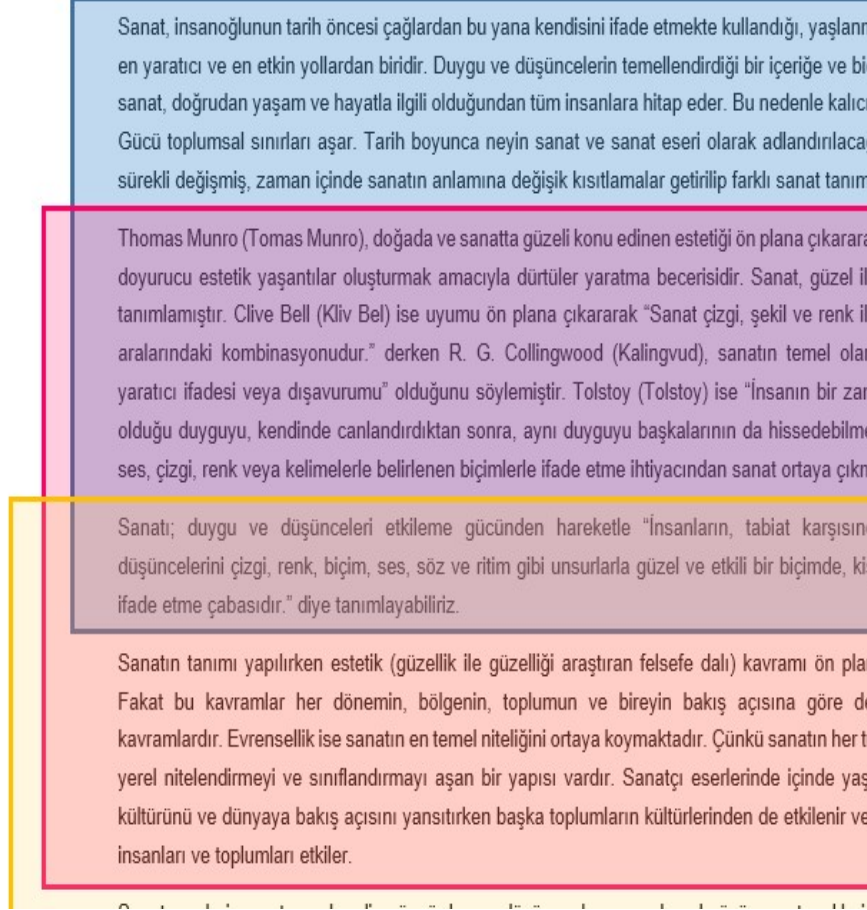
Çalışmanın ön işlem aşamasında, tüm eğitim içerikleri sözcük dizileri olarak ele alınmıştır. İlk olarak, dokümanlardaki tüm sözcükler küçük harflere dönüştürülmüş, sonrasında satır sonu işaretleri karakterleri ve alfabe dışı aralıklama karakterleri dokümanlardan ayıklanmıştır. Dokümanlar; simge, sembol, değişken ve matematiksel formüllerden ile sayısal ifadelerden arındırılmıştır. Daha sonra özel karakterler (#, %, & ve \$ gibi) ve noktalama işaretleri çıkarılmıştır. Ön işlem sonucunda, eğitim içerikleri, ardışık sözcükler arasında boşluklar bulunan bir sözcük dizisi şeklinde düzenlenmiştir.

2.2 En Uzun Gövde Tespiti

Doğal dil işleme ve bilgi geri getirim alanlarının temel sorunlarından biri olan gövdeleme, İngilizce gibi biçimsel yapısı farklı olan dillere göre daha kolay bir işlem olsa da, Türkçe gibi eklemeli dillerde ekin cümlenin anlamını ve yapısını değiştirebilen özelliği nedeniyle daha karmaşıktır. Bu nedenle, bu tezde gövdeleme işlemi üzerinde önemle durulmuştur. Çalışmada kullanılan gövdeleme yöntemi hece tabanlı bir yaklaşımdır. Bu yaklaşım, Türkçe heceleme algoritması kullanarak çalışır. Algoritma öncelikle sesli harflerin yerini belirler ve sesli harf yoksa, sözcüğü ilk sesli harften başlayarak hecelere ayırır. Yöntem, eklerin "ile" kelimesinin yazımı ve sert sessiz yumuşaması kuralını dikkate alır ve ek almış bir kelimenin en uzun gövdesini bulmaya çalışır. Bu süreçte ikili arama algoritması kullanılarak, ekler atıldığında kalan kısmın sözlükte bulunup bulunmadığı kontrol edilir. Eğer kalan bölüm sözlükte varsa, sözcüğün en uzun gövdesi olarak belirlenir.

Yöntemin devamında, önişlemden geçirilerek bölütlenmeye hazır hale getirilen belgeler için sezgisel bir yaklaşım sunulmaktadır. Öncelikle, Şekil 2.2'de gösterildiği gibi, metinler birbiriyle ortak paragraflar içeren alt bloklara ayrılır. Şekil 2.2, Millî Eğitim Bakanlığı'nca hazırlanan Sanat Tarihi kitabından alınan bir örnek üzerinden, birbiriyle çakışan alt metin bloklarının nasıl çıkarıldığını göstermektedir. Bu yaklaşım, paragraflar arasındaki ilişkileri belirlemeyi amaçlar ve çoklu etiketleme sürecini kolaylaştırır, bu bağlamda özgün bir fikirdir. Eğitim dokümanlarında yer alan iç içe geçmiş girift bir yapının (örneğin 1-3. paragraflar arası veya 25-27. paragraflar arasının) hangi öğretim kavramları ile ilişkili

olduğunun belirlenmesinin, eğitim dokümanının bir bütün olarak değerlendirilmesi bağlamında önemli bir bakış açısı sağlayacağı düşünülmektedir.



Şekil 2.2 Ortak paragraf içeren metin bloklarının ayrıştırılması

Ardından elde edilen metin blokları artık ayrı birer metin dosyası olarak algılanmış ve her metin bloğu için ayırt edici özniteliklerin tespit edilmesine çalışılmıştır. Bu da metin bloklarını temsil eden öznitelik vektörlerinin belirlenmesini güçleştiren faktörlerden biridir. Çünkü metin blokları, dokümanlara kıyasla çok daha az sözcük içermektedir. Bu sorunun üstesinden gelmek için, öznitelik vektörleri oluşturulurken en uzun gövde biçimindeki 1'li ve 2'li sözcük dizilimleri (monogram ve bigramlar) çıkartılmış ve bir sözcük torbası modeli oluşturulmuştur. Bu aşamada sözcük torbası modelinin boyutunun nasıl belirleneceği, yeni bir araştırma problemi olarak ortaya çıkmaktadır. Her bir metin bloğu farklı sayıda monogram içerdiğinden, sözcük torbasının boyutu en çok, en az sayıda monogram içeren metin bloğunun içerdiği en uzun gövde sayısı kadar olabilir. Ayrıca, sözcük torbasının boyutunun sınıflandırma veya kümeleme için kritik olabileceği unutulmamalıdır. Öznitelik çıkarma

işleminden sonra, boyut azaltma işlemi için Temel Bileşen Analizi (Principal Component Analysis) kullanılmıştır.

2.3 n -gram Çıkarımı

Ön işlem sonrasında, içerikten n 'li terimler çıkartılmıştır. Burada n 'li terim, en uzun gövde formatında n ardışık sözcükten oluşan dizilimi temsil eder. Her bir n 'li terimin metin içindeki görünme sıklığı belirlenir ve bu terimlerin öznitelikleri oluşturulur. Bu aşama için, metin içindeki n ardışık sözcüğün başlangıç ve bitiş konumları iki ayrı dizi içinde saklanır. Daha sonra, n -gramlar hızlı sıralama algoritması (quick sort) ile sözlük sırasına göre düzenlenir. Ek olarak aynı n -gramların başlangıç ve bitiş konumlarını işaret eden diziler de sıralanır. Burada sözcüğün başlangıç konumu metin içinde sözcüğün kaçınıcı karakterden itibaren başladığı ile bitiş konumu ise sözcüğün içerdiği karakter sayısına göre tanımlanır. Hızlı sıralama algoritmasının zaman karmaşıklığı, m değeri dokümandaki n -gram sayısını göstermek üzere en iyi durumda $O(m \cdot \log(m))$ ve en kötü durumda $O(m^2)$ olur. Bu sayede aynı terimler bir arada bulunur ve terimlerin başlangıç ve bitiş konumları arasındaki farklar ardışık iki terimin aynı sözcük olup olmadığının belirlenmesinde etkin rol oynar. Terimler aynı uzunlukta ise Boyer-Moore metin karşılaştırma algoritması ile (karakter karakter sağdan sola doğru tarama yaparak) ardışık iki terimin aynı olup olmadığı kontrol edilir. ℓ karakter içeren eşit uzunluklu iki metinsel ifadenin Boyer-Moore algoritması ile karşılaştırılması işlemi en iyi durumda $O(1)$ ve en kötü durumda $O(\ell)$ zaman karmaşıklığı karmaşıklığına sahiptir. Karşılaştırılan iki değer birbiriyle aynı ise frekansları 1 artırılır ve ardından gelen ardışık iki n 'li terim karşılaştırılır. Eğer metinsel ifadeler birbirinden farklı ise ilk n 'li terim sözcük havuzuna frekansı ile birlikte eklenir. Sonuç olarak, tüm dokümandaki monogram ve bigramlar dokümandaki görüntülenme sıklıklarıyla birlikte n -gram modeline göre çıkarılmıştır. Olasılıksal bir doğal dil işleme modeli olan n -gram modeline göre bir n -gramın belli bir sözcük diziliminden sonra görüntülenmesi olasılığı (2.1) eşitliğiyle hesaplanır (Manning et al, 2009; Günel ve arkadaşları, 2016).

$$P(w_i | w_{i-(n-1)} \dots w_{i-1}) = \frac{C(w_{i-(n-1)} \dots w_{i-1} w_i)}{C(w_{i-(n-1)} \dots w_{i-1})} \quad (2.1)$$

Burada w_i , bir sonraki kelimeyi temsil etmektedir. $C(w_{i-(n-1)} \dots w_{i-1})$ dizinin oluşma sayısını, $w_{i-(n-1)} \dots w_{i-1} w_i$ derlemdeki kelimelerin sıklığını belirtmektedir.

2.4 *n*-gram Filtreleme

n'li sözcük dizilimlerinin ilk veya son sözcüğünün bir durak sözcüğü olması, aday öğretim kavramları olarak belirlenirken önemli bir sorun oluşturur. Eğitim içeriklerindeki öğretim kavramları, bazen bir durak sözcükle başlayıp bitmeyebilir, bu nedenle bu tür *n*-gram'lar dikkate alınmamalıdır. Örneğin, 'koşul ifadeleri' ve 'döngüler', 'Programlama' alanında öğretim kavramlarıdır, ancak 'koşul ifadeleri ile' ve 've döngüler' gibi sözcük dizilimleri öğretim kavramları olarak kabul edilmez.

Durak sözcükler, tek başlarına anlam taşımayan ve metin bütünü içinde düşük ayırt edicilik değerine sahip olan sözcüklerdir. Bu tür sözcükler genellikle cümle içinde sadece söz dizimsel bir rol oynarlar ve anlamsal bilgi çıkarımına katkıda bulunmazlar. Bu nedenle çoğu zaman göz önüne alınmazlar. Literatürde, durak sözcüklerin çıkarılması için bir takım yaklaşımlar yer almaktadır. Geleneksel yaklaşımlardan biri, sözcükleri durak sözcüğü havuzundaki sözcüklerle karşılaştırmaktır. Zipf yasasını temel alan yaklaşımlarda, dokümandaki yüksek frekansta görülen tek sözcük olan ifadeler, ters doküman frekansı da düşükse durak sözcüğü olarak kabul edilir ve çıkarılır. Karşılıklı Bilgi (Mutual Information) tekniğinde, durak sözcüklerin çıkarılması sınıflandırma problemi olarak görülür ve karşılıklı bilgi değeri düşük çıkan sözcükler için durak sözcüğü etiketlemesi yapılır. Terim tabanlı rassal örnekleme yaklaşımında ise, durak sözcükleri rastgele metin blokları üzerinde Kullback-Leibler ayrışma ölçüsü kullanarak belirlenir. Bu tezde, bu dört yaklaşımın ikisi birleştirilerek durak sözcüklerinin tespiti sağlanmıştır. Kumova ve Karaođlan (2017) çalışmalarında, ilk veya son sözcüğü durak sözcüğü olan *n*'li terimlerin çıkarılmasından sonra yazım denetimi kısmına geçmişlerdir. Aşlıyan, Günel ve Yakhno (2007) tarafından önerilen istatistiksel bir yaklaşımda, Türkçe harf sıralamalarının doküman içinde görüntülenme olasılıkları kullanılarak hatalı yazılmış sözcükleri tespit edilir. Aşlıyan, Günel ve Yakhno (2007) çalışmasında, yaklaşık %97 doğruluk oranı elde etmiştir ve üçlü harf sıralamalarının ardışık harf görüntülenmesi olasılıklarını kullanarak hatalı yazılmış sözcükleri tespit etmiştir. Bu tezde de, yazım denetimi uygulanmıştır ve 'yazım denetimini geçemeyen *n*-gramlar, aday öğretim kavramı olamaz' hipoteziyle yöntem geliştirilmiştir.

2.5 Öznitelik Vektörlerinin Oluşturulması

Tez içinde, öğretim kavramlarının çıkarılması için kullanılan ve dönüştürülen öznitelikler aşağıdaki gibidir:

- n : aday öğretim kavramında bulunan sözcük sayısı, $t_j = w_1 w_2 \dots w_n$
- Frekans (Frequency): aday öğretim kavramı olan t_j 'nin eğitim dokümanında görüntülenme sıklığı, $f(t_j)$
- Terim Frekansı (Term Frequency): t_j sözcük dizilimi sıklığının eğitim dokümanında yer alan tüm n -gramlarının toplam sıklıklarına oranı, $tf(t_j)$
- Metin Bloğu Frekansı (Text Block Frequency): t_j sözcük dizilimini içeren metin bloklarının sayısı, $df(t_j)$.
- *Ters Metin Bloğu Frekansı (Inverse Text Block Frequency)*: Bu çalışmada tek bir doküman üzerinde çalışıldığından, metin madenciliği yaklaşımlarında sıklıkla kullanılan özniteliklerden biri olan ters doküman frekansı değeri, Ters Metin Bloğu Frekansı olarak yeniden yorumlanmıştır. Ters Metin Bloğu Frekansı, N , d dokümanı içindeki tüm metin bloklarının sayısı olmak üzere $itbf(t_j) = \log\left(\frac{N}{df(t_j)}\right)$ eşitliği ile hesaplanır.
- *Metin Bloğu Frekansı – Ters Metin Bloğu Frekansı (Text Block Frequency – Inverse Text Block Frequency)*: Ters Metin Bloğu Frekansı özniteliğine benzer yaklaşım terim frekansı – ters doküman frekansı özniteliği için kullanılarak metin bloğu frekansı – ters metin bloğu frekansı değerleri her bir t_j sözcük dizilimi için $itbf - itbf(t_j) = tbf(t_j) \times itbf(t_j)$ olarak hesaplanır.

2.6 Normalizasyon

Tezde, aday öğretim kavramları için oluşturulan öznitelik vektörleri, min-max normalizasyon yöntemi ile standartlaştırılmıştır. Min-max normalizasyonu bir veri kümesinde yer alan tüm verilerin, bu kümedeki en büyük ve en küçük veriye göre normalleştirilmesi işlemidir. Min-max normalizasyonu sonucunda en küçük verinin değeri 0, en büyük verinin değeri 1 olmaktadır. Aradaki diğer tüm veriler ise bu aralıkta değerler alır. Bu normalizasyon

yöntemi kullanılarak çalışmadaki verilerin öznitelik değerlerinin [0,1] aralığında olması sağlanmıştır. Min- max normalizasyon yöntemine ait eşitlik aşağıda verilmiştir:

$$x' = \frac{x-min}{min-max} \quad (2.2)$$

Bu aşamada metin bloklarındaki aday öğretim kavramı sayılarının minimumu hesaplanır ve hesaplanan sayıda sözcük diziliminin öznitelik vektörleri birleştirilerek metin bloklarının öznitelik vektörleri oluşturulur. Aslında metin bloklarını betimlemede yapılan tipik bir sözcük torbası tanımlamasıdır.

2.7 Boyut İndirgeme

Normalizasyon işlemlerinden geçirilen öznitelik vektörleri, ardından temel bileşenler analizi (PCA) kullanılarak boyut indirgeme işlemine tabi tutulmuştur. Boyut indirgeme yöntemlerindeki temel amaç; verileri mümkün olan en az düzeyde özellik kaybı ile sıkıştırmak, küçültmektir. Genellikle büyük çok değişkenli veri kümeleri analiz edildiğinde boyutlarını azaltma ihtiyacı doğar ve bu amaçla temel bileşenler analizi, yaygın olarak kullanılan önemli bir teknik olarak görülür (Jolliffe, 2005). Bu yöntem ilk olarak Pearson (1901) tarafından ortaya koyulmuştur. PCA, en popüler istatistiksel tekniklerden biridir ve birçok bilimsel disiplin tarafından kullanılır (Abdi ve Williams, 2010). Bir dizi muhtemel ilişkili değişkeni doğrusal ilişkisiz bir dizi değişkene dönüştürür (Günel ve arkadaşları, 2016). PCA, ancak değişkenler arasında güçlü bir korelasyon olduğunda etkili bir yöntemdir (Kim, 2009). Yöntemin *temel bileşenler* olarak adlandırılmasının nedeni, veri kümesindeki veri hakkında en çok bilgi içeren bileşenlerin kullanılıyor olmasından (en küçük karesel hatanın hesaplanmasından) ileri gelmektedir. Özetle bu yöntem, eldeki veri kümesine ait en temel özelliklerin yakalanması mantığına dayanmaktadır. Bu yöntemi gerçek dünyadan bir benzetimle açıklamak gerekirse, bir buzdolabındaki yiyecekler veri kümesi olarak düşünülebilir. Yiyecekleri belli özelliklerine göre yerleştirerek buzdolabındaki alanı en verimli şekilde kullanmak amaçlandığında, yiyecekleri türlerine göre (kahvaltılıklar, süt ürünleri, sebzeler, vs) sınıflandırmak, renklerine göre (kırmızı yiyecekler, yeşil yiyecekler vs) sınıflandırmaktan daha mantıklı bir yaklaşım olacaktır. Bu durumda yiyeceklerin türleri, renklerinden daha önemli (daha temel) bir bileşen olacaktır. Problem durumu değiştiğinde, yiyeceklerin farklı özellikleri daha önemli (daha temel) bir hâle gelebilir.

Temel bileşenler analizi yöntemini matematiksel olarak açıklamak gerekirse, bir veri kümesinin en küçük ortalama karesel hata ile daha küçük bir alt uzayda ifade edilmesini sağlayan dönüşüm matrisi, verinin kovaryans matrisinin özvektörleridir denebilir (Abacı, 2019). Aşağıdaki P matrisinin temel bileşenleri (özvektörleri), eşitlik 2.3'deki gibi hesaplanır.

$$\mathbf{P} = [p_1, p_2, \dots, p_D] \text{ ve } \boldsymbol{\lambda} = [\lambda_1, \lambda_2, \dots, \lambda_D], \mathbf{X} \in \mathbb{R}^D \times \mathbb{N}$$

$$\frac{\partial L(p, \lambda)}{\partial p} = 2C_p - 2\lambda_p = 0 \Rightarrow C_p = \lambda_p \quad (2.3)$$

Çalışmanın ilerleyen aşamasında, belirli bir alandan bağımsız olarak etiketleme yapma amacı güdüldüğünden denetimsiz kümeleme yaklaşımının kullanılması uygun görülmüştür.

Bundan sonraki aşamada, öncelikle K-Means kümeleme yöntemi ve ardından Parçacık Sürü Optimizasyon (PSO) yaklaşımı açıklanacaktır. Daha sonra, bu yaklaşımların çalışmada nasıl kullanıldıkları anlatılacaktır.

2.8 Kümeleme

Çalışmada, etiketleme sürecini gözetimsiz bir şekilde otomatikleştirmek hedeflenmiştir. Bu nedenle sınıflandırma yaklaşımı yerine kümeleme yaklaşımı tercih edilmiştir. Bu tercih, Yann Le Cunn'un denetimli ve denetimsiz öğrenmeyi anlatan ve oldukça ünlü şu cümlesi ile açıklanabilir: "Zekâ bir pasta olsaydı, denetimsiz öğrenme pastanın keki olurdu, denetimli öğrenme pastanın kreması olurdu ve pekiştirmeli öğrenme pastadaki kiraz olurdu." (Geron, 2019). Bu cümleden de anlaşılacağı üzere, makine öğrenmesinin temelini oluşturan yaklaşımlar kümeleme yaklaşımlarıdır. Su ve arkadaşları (2010), kümelemenin birçok disiplinde sayısız uygulaması olan temel bir veri analizi uygulaması olduğunu belirtmiş ve kümelemeyi, aynı kümenin öğelerinin birbirine daha fazla benzemesi için verileri gruplara ayırma işlemi olarak tanımlamıştır.

2.8.1 K-Means: K-means algoritması, kümeleme problemlerini çözmek için kullanılan denetimsiz bir öğrenme algoritmasıdır (MacQueen, 1967). Temel mantığı, n adet veriden oluşan bir veri kümesini, k adet kümeye ayırma amacına dayanmaktadır (Sarıman, 2011). Diğer kümeleme algoritmalarında olduğu gibi hedef, küme içindeki benzerliklerin en

yüksek ve kümeler arasındaki benzerliklerin en düşük olmasının sağlanmasıdır. Sinaga ve Yang (2020), K-means algoritmasında küme sonuçlarının başlangıç noktasından önemli ölçüde etkilendiğini belirtmişlerdir.

K-Means yöntemi, Öklid uzaklığı formülü temel alınarak geliştirilmiştir:

$$p = (p_1, p_2, \dots, p_n) \text{ ve } q = (q_1, q_2, \dots, q_n)$$

$$\sqrt{\sum_{i=1}^n (p_i - q_i)^2} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} \quad (2.4)$$

K-Means algoritmasının adımları şu şekildedir: Rastgele seçilen k adet küme merkezi ile başlanır. $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_k$ aşağıdaki gibi hesaplanır:

$$\mathbf{M}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ik} \quad (2.5)$$

Ardından karesel hata formülü ile küme içi değişimler hesaplanır ve toplam karesel hata bulunur:

$$e_i^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2 \quad (2.6)$$

$$E_k^2 = \sum_{k=1}^K e_k^2 \quad (2.7)$$

Bu işlemlerden sonra veri grubundaki her bir veri, kendine en yakın küme merkezinin kümesine atanır. Daha sonra yeni küme merkezleri, içindeki verilerin ortalaması alınarak hesaplanır. Küme merkezlerinde bir değişiklik olmayıncaya kadar bu atama ve hesaplama işlemi devam eder.

2.8.2 Parçacık Sürü Optimizasyonu: Çalışmanın bu bölümünde öncelikle sezgi, sezgi üstü ve sürü zekası kavramlarına kısaca değinilecek ve ardından PSO yaklaşımı anlatılacaktır.

Sezgi (Heuristic) ve Sezgi Üstü (Metaheuristic): Sezgi, Türk Dil Kurumu (2022) tarafından ‘sezme yeteneği, feraset’ olarak tanımlanmıştır. Sezgisel ve sezgi üstü algoritmalar doğadaki canlıların ‘sezme’ yeteneklerinden, doğada gerçekleşen bir takım olay ve olgulardan esinlenilerek ortaya çıkmış algoritmalarıdır. Bu algoritmalar, çözüm uzayında kesin çözümü değil en yakın çözümü bulurlar. Kesin çözüme ulaştırmıyor olmalarına rağmen sezgisel ve

sezgi üstü algoritmalarla neden ihtiyaç duyulduğunu Karaboğa (2011), optimizasyon problemlerinin kesin çözüme yakınsama amaçlı olmasıyla ve karar verme işleminin de öğrenme sürecinin bir parçası olarak görülmesiyle açıklamıştır. Yaşamdaki problemlerin çözümünde de bir miktar ‘sezgisellik’ bulunmakta, elde edilen sonucun bir diğer yöntem seçildiğinde elde edilen sonuçtan daha iyi (doğru) olduğuna dair bir kesinlikten söz edilememektedir. Sezgiler, tarih boyunca temel bilimlerden felsefeye, sanattan edebiyata insanlığın gelişimde yer alan bir kavramdır. Bu bağlamda düşünüldüğünde, ‘kesinlik’ kavramını içermeyen sezgisel ve sezgi üstü algoritmaların bilgisayar bilimlerinde neden sıklıkla tercih edildiği daha iyi anlaşılabilir.

Sürü Zekası (Swarm Intelligence): Tüm canlı davranışlarının en temelinde hayatta kalma güdüsü vardır. Bazı canlıların tek başlarına ortaya koydukları davranışlar hayatta kalmaları için yeterli değildir. Sürü zekası, canlıların ortak bir şekilde bir takım davranışları sergileme yetenekleri olarak tanımlanabilir. Kuşların sürü davranışlarını ayrıntılı şekilde ele alıp modelleyen ve oldukça popüler olan çalışmada Reynolds (1987), kuş sürülerinin ve diğer hayvan sürülerinin senkronize hallerini izlemenin güzel, keyifli olduğunu ve üzerinde düşünmeye değer olduğunu söylemiştir. Bir canlı topluluğundaki tüm bireyler topluluktaki en iyi, en faydalı davranışı gösteren bireylerin davranışları ve kendi deneyimlerini birleştirerek bir çıkarım yaparlar. Bu çıkarımlarını, karşılaştıkları problemleri çözmek için kullanırlar ve böylece bu topluluğa özel bir takım davranış kalıpları ortaya çıkar. Canlıların bu toplu olarak sergiledikleri davranışlar incelenerek sürü zekası anlayışını temel alan optimizasyon algoritmaları ortaya koyulmuştur. Birçok farklı sürü zekası algoritması vardır. Bir örnek vermek gerekirse, ateş böceği algoritmasından bahsedilebilir. Ateş böceklerinin ışıklarını yakıp söndürmesindeki amaç arkadaşlarını bulmak, kendi avını kendisine çekmek ve avcılarından korunmaktır. Bir ateş böceği, kendisinden daha parlak ışık yakan başka bir ateş böceği gördüğünde ona doğru çekilir. Bu iterasyon bu şekilde devam eder ve sürüdeki tüm ateş böcekleri, en parlak ışığı olan ateş böceğinin etrafında toplanırlar. İşte bu davranış, Yang (2007) tarafından geliştirilen ateş böceği algoritmasının temel mantığını oluşturur. Sürü zekası algoritmaları sezgisel ve sezgi üstü yöntemlerdir.

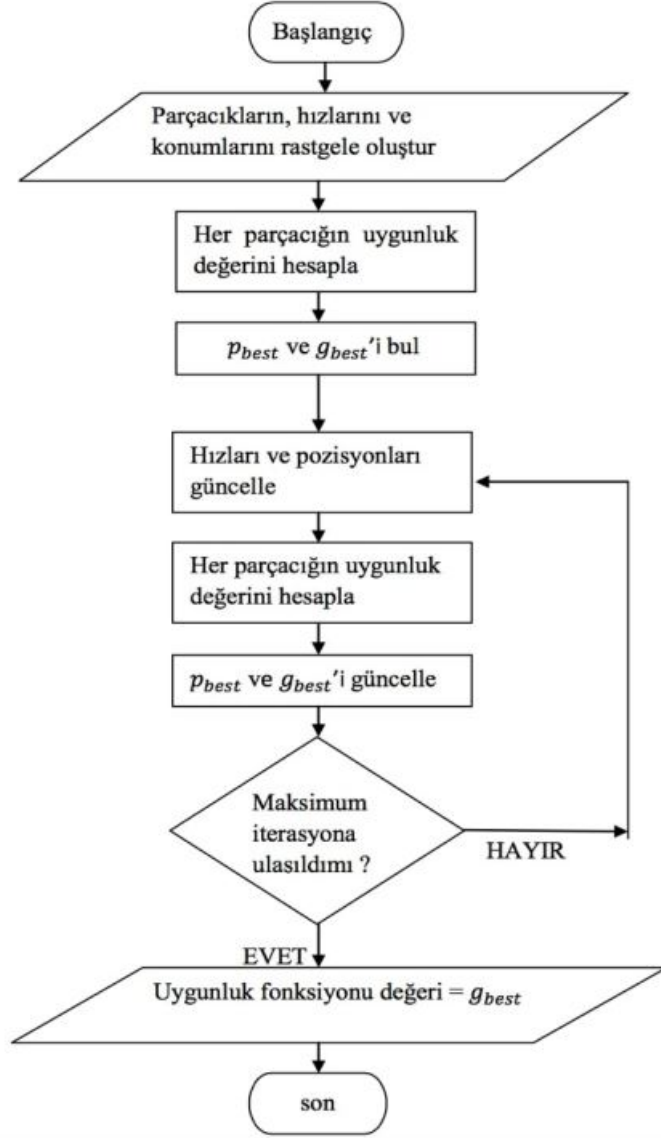
PSO: Optimizasyon, bir şeyi en iyi hale getirmek ya da en uygun şekilde kullanmak anlamına gelir (TDK, 2022) ve insanlık tarihi boyunca üzerinde düşünülen ve çalışılan bir kavramdır. PSO, Kenedy ve Aberhart (1995) tarafından ortaya koyulan sezgi üstü bir optimizasyon yöntemidir. Sürü zekası kavramına dayanmaktadır ve kuşlar, karıncalar gibi sürü hâlinde sosyal davranışlar sergileyen hayvanlardan esinlenilerek ortaya koyulmuş bir

algoritmaya sahiptir. Kennedy ve Aberhart (1995), balık ve kuş sürüleri gibi hayvanların davranışlarını modellemek için çalışırken, ortaya koydukları algoritmanın başarılı bir şekilde optimizasyon testleri yapabildiğini de görmüşler ve PSO adını verdikleri optimizasyon yöntemini ortaya koymuşlardır (Heris, 2015). Oldukça güçlü olan PSO algoritması fen, mühendislik, sosyal bilimler gibi birçok alandaki büyük boyutlu verilerin optimizasyonunda sıklıkla ve başarıyla kullanılmaktadır.

Parçacık Sürü Optimizasyonu (PSO), çok yönlü ve popülasyon temelli bir stokastik (rastgele) optimizasyon yöntemidir (Yang ve Le, 2010). PSO yaklaşımı kullanılırken, başlangıçta arama uzayında bulunan tüm parçacıklar rastgele bir şekilde yerleştirilir. Parçacıkların uygunluk değerleri, optimize edilecek fonksiyonun parçacığın bulunduğu konumundaki değerleri olarak belirlenir (Gör, 2020). Her bir parçacık için hız ve konum bilgisi oluşur, ardından yine sürüdeki parçacıkların uygunluk değerleri bulunur. Sürü içinde parçacıklar, tüm sürü zekası algoritmalarında olduğu gibi kendi bilişsel bilgileri ile sürü davranışlarından öğrenerek kazandıkları bilgiyi bir arada kullanırlar. Her bir parçacığın hız, konum ve uygunluk değeri bilgileri, her bir iterasyonda güncellenir ve her bir parçacık, en iyi konuma sahip parçacığın etrafında hızlı bir şekilde toplanırlar. PSO algoritmasında sürü ifadesi, olası çözüm kümesini gösterir (Merwe and Engelbrecht, 2003) ve aşağıdaki eşitlikle gösterilir:

$$P = \{p_1, p_2, \dots, p_s\} \quad (2.8)$$

Her bir parçacık, olası konumları tarayarak kendinden daha iyi olan konuma doğru yaklaşır. Her bir adımda (t) her bir parçacık (p_i) kendi hız (\mathbf{V}_i) ve konum (\mathbf{x}_i) bilgileri ile elde ettiği en iyi amaç fonksiyonu değerini veren konum bilgisini ($\mathbf{p}_{i,best}$) hafızasında tutar. Yeni adımlarında bu bilgileri kullanarak tekrar en iyi konumu aramaya devam eder. Ayrıca parçacığın $N_i \subseteq P$ yakın komşuluğundaki diğer parçacıklara ait $\mathbf{p}_{j,best}$ verileri de karar sürecinde dikkate alınır. Yani parçacıklar arasında her bir adımda bilgi alışverişi söz konusudur. Bununla birlikte parçacığın hareket doğrultusunu belirleyen hız vektörünün hesaplanmasında sürüdeki o ana kadar bilinen en iyi çözümün konum bilgisi de kullanılır. Bu doğrultuda \mathbf{g}_{best} vektörü sürünün liderinin konumu belirtir. Böylece bilişsel olarak sürünün tamamının ürettiği bilgi kullanılmış olur. PSO algoritmasının çalışma prensibi özetle Şekil 2.3’de gösterilmiştir.



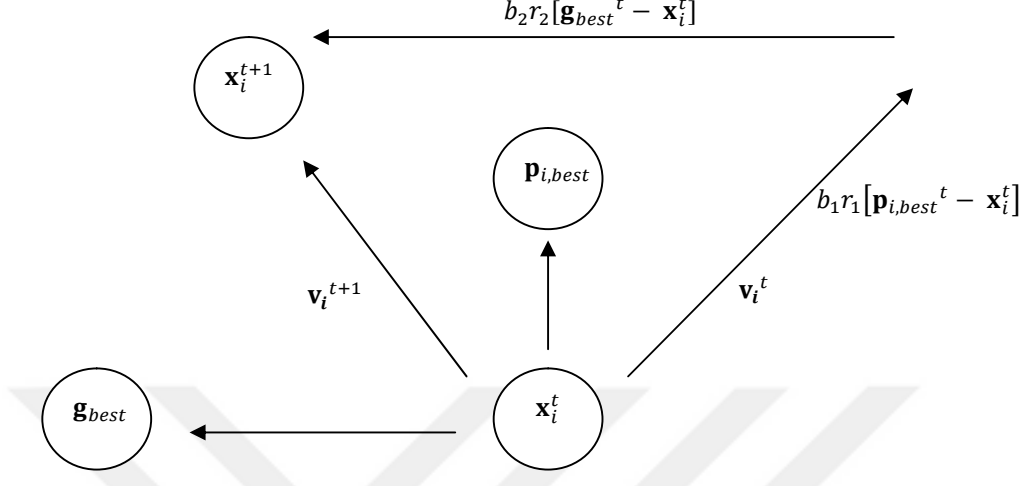
Şekil 2.3 PSO algoritmasının akış şeması (Gör'den 2020)

PSO'nun temeli, her bir iterasyonda parçacıkların hız bilgilerinin güncellenmesi mantığına dayanır. Her bir parçacığın yeni konumu, güncellenen hız bilgisine göre yeniden güncellenir (Pehlivanoglu, 2017). Bir parçacığın herhangi bir adımdaki hızı ve konumu aşağıdaki eşitlikler ile belirlenir.

$$\mathbf{v}_i^{t+1} = a\mathbf{v}_i^t + b_1r_1[\mathbf{p}_{i,best}^t - \mathbf{x}_i^t] + b_2r_2[\mathbf{g}_{best}^t - \mathbf{x}_i^t] \quad (2.9)$$

$$\mathbf{x}_i^{t+1} = c\mathbf{x}_i^t + d\mathbf{v}_i^{t+1} \quad (2.10)$$

Şekil 2.4'de parçacığın bir sonraki konumuna yönelimi gösterilmiştir.



Şekil 2.4 Parçacıkların her bir iterasyonda yeni konumuna yönelimi

Bu tezde PSO yaklaşımının kullanılmasının nedeni, metin bloklarının otomatik olarak etiketlenmesinin alandan bağımsız olarak hedeflenmiş olmasıdır.

3. DENEYSEL ÇALIŞMALAR

Sanat Tarihi, Türk Kültür ve Medeniyet Tarihi kitaplarının kaynak olarak kullanıldığı deneysel çalışmalarda, öğretim kavramlarının öncelik derecelerine göre otomatik olarak eğitim içeriklerinden tespit edilmesi ve metnin tespit edilen öğretim kavramlarıyla etiketlenmiş alt metin bloklarına bölütlenmesi hedeflenmiştir. Paragrafların birden çok öğretim kavramı ile etiketlenmesi de çalışmanın amaçlarından biridir.

İlk olarak ön işlemden geçirilen tüm dokümanlar, çakışan metin blokları şeklinde ayrılmıştır. Önışlemin ardından hece tabanlı gövdeleme yaklaşımı uygulanmıştır. Hecelemenin ardından Türkçe diline has kurallar bu aşamada uygulanmıştır. Sözcüklerin hecelenmesi ve kelimenin sonundan başına doğru hecelerin atılarak en uzun gövdenin bulunması sırasında dikkate alınan kurallar Tablo 3.1’de verilmiştir.

Tablo 3.1 En uzun gövde tespiti aşamasında uygulanan işlemler ve açıklamaları

İşlem	Açıklaması
Bulunma hali	İsmin -de, -da, -te veya -ta eklerinden birini aldığı durumdur.
Belirtme hali	İsmin -ı, -i, -u ve -ü eklerinden birini aldığı durumdur.
Ayrılma hali	İsmin -den, -dan, -ten veya -tan eklerinden birini aldığı durumdur.
Yönelme hali	İsmin -e veya -a eklerinden birini aldığı durumdur.
İyelik eki	İsimlerin sonuna gelerek onlara aitlik anlamı katar.
İle'nin yazılışı	İle sözcüğü, ünsüzle biten sözcüklerle bitişik olarak yazıldığında i ünlüsü düşer ve büyük ünlü uyumuna uyar. Ünlüyle biten sözcüklerle bitişik olarak yazıldığında araya y ünsüzü girer ve başındaki i ünlüsü düşer

En uzun gövde belirlenmesinin ardından, n -gram çıkarımı yapılmıştır.

Sanat Tarihi kitabında “Estetik”, “Heykel”, “Mimari”, “Resim”, “Sanat”, “Sanatçı”, “Zanaatkâr”, “Tarih”, “Tunç”, “Tablet”, “Karum”, “Tümülüs”, “Megaron” ve “Sikke” sözcükleri olmak üzere on dört öğretim kavramı verilmiştir. Türk Kültür ve Medeniyet Tarihi kitabında ise “Bağımsızlık”, “Kut”, “Toy”, “Töre”, “Ülke”, “Divan”, “Saray”, “Reâya”,

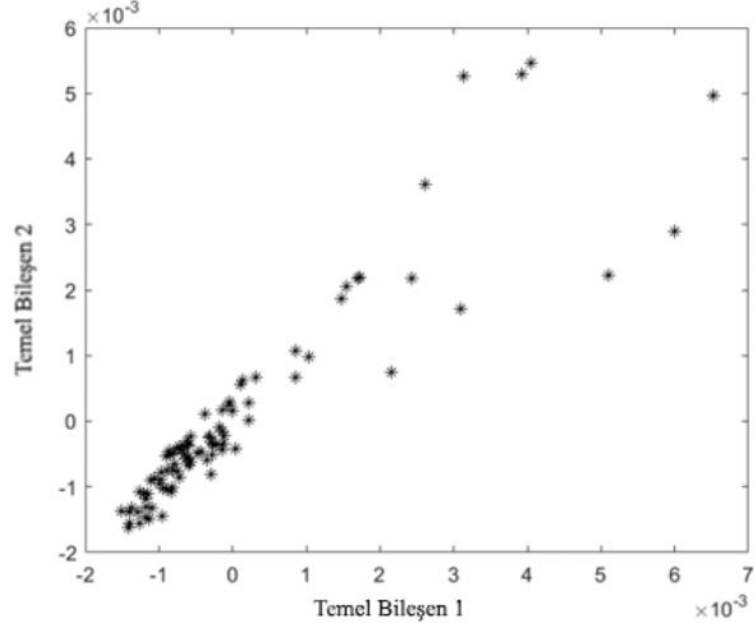
“İstimalet”, “Ulus” olmak üzere on öğretim kavramı verilmiştir. Örnek teşkil etmesi için, Sanat Tarihi kitabı ile ilgili bazı istatistiki bilgiler Tablo 3.2’de verilmiştir (Kayahan ve arkadaşları, 2022).

Tablo 3.2 Doküman hakkında bazı istatistiki bilgiler (Sanat Tarihi)

Nitelik	Değer
Dokümandaki toplam sözcük sayısı	6213
Durak sözcüklerin sayısı	1196
Paragraf sayısı	91
1-li en uzun gövde sayısı	5017
Teklikle belirlenmiş 1-li en uzun gövde sayısı	1546
1-li en uzun gövde sayısı	5351
Teklikle belirlenmiş 2-li en uzun gövde sayısı	4843
Metin bloğu başına ortalama 1-li en uzun gövde sayısı	369,856
Metin bloğu başına ortalama teklikle belirlenmiş 1-li en uzun gövde sayısı	138,747
Paragraf başına ortalama 1-li en uzun gövde sayısı	68,659
Paragraf başına ortalama teklikle belirlenmiş 1-li en uzun gövde sayısı	55,131
Öğretim kavramı sayısı	14
Öğretim kavramları sayısının teklikle belirlenmiş 1-li en uzun gövde sayısı oranı	0,009
Uzman tarafından etiketlenen ve paragraf başına ortalama öğretim kavramı sayısı	1,626
Uzman tarafından tek öğretim kavramıyla etiketlenen metin bloğu sayısı	49
Uzman tarafından çift öğretim kavramıyla etiketlenen metin bloğu sayısı	27
Uzman tarafından üç öğretim kavramıyla etiketlenen metin bloğu sayısı	15
Sistem tarafından etiketlenen ve paragraf başına ortalama öğretim kavramı sayısı	2,286
Uzman tarafından tek öğretim kavramıyla etiketlenen metin bloğu sayısı	12
Uzman tarafından çift öğretim kavramıyla etiketlenen metin bloğu sayısı	41
Uzman tarafından üç öğretim kavramıyla etiketlenen metin bloğu sayısı	38

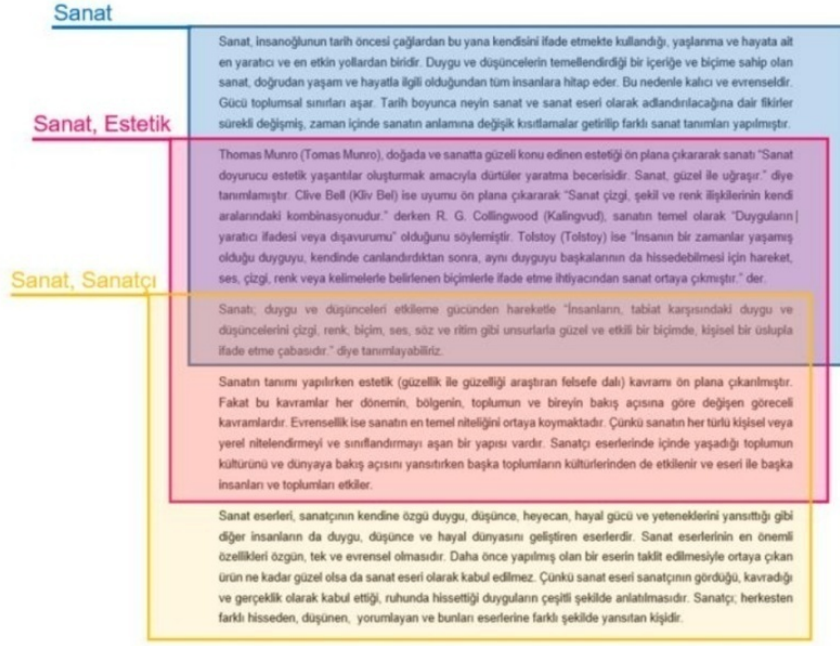
Sanat Tarihi kitabında 91 metin bloğu oluşmuş ve alan uzmanından, metin bloklarını bahsi geçen öğretim kavramları ile etiketlenmesi istenmiştir. Tablo 3.2’de de görüldüğü gibi, manuel etiketlemenin ardından, her bir metin bloğu 14 kavram üzerinden ortalama 1,626 öğretim kavramı ile etiketlenmiştir. 91 metin bloğundan yalnızca 15 tanesi 3 farklı öğretim kavramı ile etiketlenmiştir. Uzman, 27 metin bloğunu ise 2 ayrı öğretim kavramı ile etiketlemiştir. Kalan 49 metin bloğu ise 1 öğretim kavramı ile etiketlenmiştir.

Tüm doküman için n-gram çıkarımının ardından, sırasıyla öznitelik vektörlerinin oluşturulması ve ardından Temel Bileşen Analizi ile boyut indirgeme işlemleri uygulanmıştır. Sanat Tarihi kitabına uygulanan PCA işleminden sonra metin bloklarının düzlemdeki hali Şekil 3.1’de gösterilmiştir.



Şekil 3.1 Sanat Tarihi ders kitabı için PCA işleminden sonra metin bloklarının düzlemde konumlandırılması

Çalışmanın amacı metin bloklarını (paragrafları) bağımsız olarak otomatik olarak etiketlemektir. Bu nedenle Parçacık Sürü Optimizasyonu yöntemi tercih edilmiştir. Parçacık Sürü Optimizasyonu, denetimsiz öğrenme kullanarak kümeleme yapmak için kullanılmıştır. Kümeleme sonucunda metin blokları, Şekil 3.2’de gösterilen bir bölümde olduğu gibi her bir paragraf en az bir en fazla üç kavramla etiketlenmiştir.



Şekil 3.2 Ortak paragraf içeren metin bloklarının etiketlendirilmesi (Sanat Tarihi kitabı)

Oluşturulan sistemin performans ölçümü, sistem çıktıları ile uzman görüşlerinin karşılaştırılmasıyla yapılmıştır. Performans ölçütü için izlenen yol kısaca aşağıdaki gibidir:

LC kümesi, N paragraf içeren d dokümanından elde edilen öğretim kavramlarının kümesi olsun. $k = 1, 2, \dots, N$ için,

$$T_k = \{t_{k,i} \in LC \mid i \in \{1, 2, 3\} \text{ için } t_{k,i} \text{ değeri } k. \text{ paragraf için alınan uzmanının belirlediği etiket}\}$$

$$O_k = \{o_{k,j} \in LC \mid j \in \{1, 2, 3\} \text{ için } o_{k,j} \text{ değeri } k. \text{ paragraf için sistemin belirlediği etiket}\}$$

küme için $i, j \in \{1, 2, 3\}$ olacak şekilde

$$\chi_k(i, j) = \begin{cases} 0, & o_{k,j} = t_{k,i} \text{ ise} \\ 1, & \text{aksi halde} \end{cases}$$

fonksiyonu tanımlansın. Bu durumda çoklu etiketleme modeli için ortalama karesel hata (Mean Squared Error, MSE)

$$MSE = \frac{1}{N} \sum_{k=1}^N (\chi_k(i, j))^2$$

eşitliği ile hesaplanır. Bu temel yaklaşıma ek olarak, etiketler için önem sıralarına göre sezgisel bir ağırlıklandırma yaklaşımı uygulanmıştır: Her paragrafın en az bir en çok üç kavramla etiketlendirildiği göz önünde bulundurularak, önem derecesine göre hedef etiketler sırasıyla $i = 1, 2, 3$ ile indislendirilmiş ve ağırlıklar sırayla 0,6, 0,3 ve 0,1 değerleriyle derecelendirilmiştir. Bu yaklaşıma göre ağırlıklandırılmış ortalama karesel hata miktarı (Weighted Mean Squared Error, wMSE)

$$\text{wMSE} = \frac{1}{N} \sum_{k=1}^N (0,6\chi_k(1,j)^2 + 0,3\chi_k(2,j)^2 + 0,1\chi_k(3,j)^2)$$

eşitliği ile hesaplanır. Bu eşitlikte $i, j = 1, 2, 3$ için $\chi_k(i, j)$ yine aşağıdaki gibi tanımlanır ve özetle, sistemin kaç adet terimi yanlış olarak etiketlediğini tespit eder (Kayahan ve arkadaşları, 2023b).

$$\chi_k(i, j) = \begin{cases} 0, & o_{k,j} = t_{k,i} \text{ ise} \\ 1, & \text{aksi halde} \end{cases}$$

Burada $o_{k,j}$ değeri alan uzmanı tarafından etiketlenen j . öğretim kavramını, $o_{k,i}$ değeri ise metnin k . metin bloğu için çıktı olarak sistem tarafından belirlenen i . terimi ifade eder.

Çalışmada, tüm metin ilk paragraftan başlayarak üçerli gruplara ayrılmıştır. Böylece tüm metin 1-3., 2-4., 3-5. paragrafları kapsayacak şekilde çakışan metin blokları sezgisel olarak oluşmuştur. Ayrıca dokümanın giriş ve sonuç bölümlerinde birbirine benzer konulardan söz edilebileceği fikrinden yola çıkılarak, son paragrafta gelindiğinde tekrar başa dönülerek metin bölütleme yaklaşımına devrinsel bir nitelik kazandırılmıştır. Bu stratejinin ana mantığı, bir dokümanın başlık ve sonuç bölümlerinde benzer kavramların geçebileceği düşüncesine dayanmaktadır. Bu nedenle oluşturulan son metin blokları, örneğin Sanat Tarihi kitabı için, 90, 91 ve 1. paragrafları ile 91, 1 ve 2. paragrafları içermektedir.

4. SONUÇ VE TARTIŞMA

Çalışma sonucu, Sanat Tarihi kitabı için ortalama karesel hatalar k-means için $wMSE = 0,557$ ve PSO kümeleme yaklaşımı için ise $wMSE = 0,527$ olarak elde edilmiştir. Türk Kültür ve Medeniyet Tarihi kitabında ise k-means için $wMSE = 0,522$ ve PSO için $wMSE = 0,531$ şeklindedir. Bu sonuçlar, her bir metin bloğu için ağırlıklandırılan en önemli öğretim kavramının net olarak belirlendiği, ikincil derecede önemli öğretim kavramının da bir kısmının doğru şekilde belirlendiği biçiminde yorumlanabilir. Ağırlıklandırılmış ortalama karesel hata formülünden anlaşılacağı üzere bir paragrafa dair yapılan etiketlemelerin hepsi hatalıysa, yani bu paragrafa ait hiçbir öğretim kavramı tespit edilemezse, 1 birimlik bir hata elde edilir. Yani N paragraf için en fazla N birimlik bir hata elde edilebilir. Eğer her paragraf kesin olarak üç farklı öğretim kavramı ile etiketlenirse böyle bir maksimum hata miktarı görülür. Ancak, bazı paragraflar daha az sayıda öğretim kavramı ile etiketlenmişse, etiket ağırlıkları güncellenerek maksimum hata miktarının hala N olmasını sağlamak mümkündür. Bu amaçla, öğretim kavramlarının metin blokları ile olan ilişkilerini derecelendiren ağırlıkların toplamının 1 olması yeterlidir.

Paragraf başına düşen ortalama hata miktarı yine 1 birimdir. Bu nedenle, hata $wMSE$ değeri $[0,1]$ aralığında olmalıdır. Başka bir deyişle, çalışmada elde edilen 0,5 birimlik ağırlıklandırılmış ortalama karesel hata, her paragraf için üç etiketten birinin kesin olarak tespit edilebildiğini, ikinci etiketin ise bir bölümünün doğru olarak belirlenebildiğini göstermektedir.

Çalışmada Sanat Tarihi, Türk Kültür ve Medeniyet Tarihi olmak üzere iki adet kitap kullanılmıştır. Sanat Tarihi kitabının yapısı incelendiğinde kullanılan kavramların, cümle ve sözcüklerin bugünün Türkçe'sine yakın ve uygun olduğu açıkça gözlemlenebilmektedir. Ayrıca paragrafların ilişkilendirildiği öğretim kavramlarının ilgili paragraflarda sözcük olarak yer aldığı görülmektedir. Örneğin sanat kavramı ile etiketlenen bir paragrafta sanat sözcüğü bazen bir, bazen birden fazla sayıda kullanılmıştır. Türk Kültür ve Medeniyet Tarihi kitabında ise, yapı biraz daha farklıdır. Kullanılan kavramlar, genellikle artık geçerliliği olmayan, anlamları az bilinen sözcüklerdir. Örneğin bu kitapta geçen 'reaya' kavramının anlamı Türk Dil Kurumu sözlüğünde (2023) 'bir hükümdarın yönetimi altındaki halk' olarak tanımlanmıştır. Bu sözcük, bugün sık kullanılmamaktadır. Ayrıca bu kavramın etiklendiği paragraflarda reaya sözcüğü neredeyse hiç geçmemekte, anlam olarak yer almaktadır.

Görüldüğü üzere çalışmada kullanılan eğitim dokümanları birbirinden farklı yapıdadır. Çalışmada kullanılan yöntem, iki farklı dokümanda birbirine yakın sonuçlar vermiştir, dolayısıyla tutarlı bir yöntem olarak yorumlanabilir.

Çalışmanın sonuçları yorumlanırken dikkate alınması gereken bir diğer önemli husus, literatürde yer alan çoklu etiketleme yaklaşımlarının başarı oranlarıdır. Çalışmanın Giriş bölümünde geniş bir şekilde ele alınan yaklaşımlarda, başarı oranlarının tekli etiketleme yaklaşımlarına oranla düşük olduğu ve halen yeterli olgunluğa ulaşamadığı görülmüştür. Bu bağlamda ele alındığında, elde edilen sonuçların oldukça başarılı olduğu çıkarımı yapılabilir. Ek olarak, çoklu etiketleme yaklaşımlarının eğitim alanında uygulandığı bir çalışmaya literatürde rastlanmamıştır. Eğitim teknolojileri penceresinden bakıldığında, yürütülen çalışmanın literatüre katkı sağlayan özgün bir çalışma olduğu görülmektedir.

Yine de sonuçların iyileştirilmesi için bir takım çalışmalar yürütülebilir. Örneğin kavramın ilgili paragrafta hiç yer almadığı durumlarda eş anlamlı sözcükler, kavram haritaları gibi sözcükler arası ilişkileri gösteren yapılar kullanılabilir. Günel (2006) kavram haritası modelinin konu içeriklerinin genel bir yapısını sunduğunu belirtmiştir. Kavram haritaları ve TreeBank gibi yapılar kullanılarak farklı sonuçlar elde edilebilir.

KAYNAKLAR DİZİNİ

- Abacı, B.**, 2019, Temel Bileşen Analizi, <https://imlab.io/2019/09/01/principal-component-analysis/> Erişim tarihi: 12/11/2022.
- Abdi, H. and Williams, L. J.**, 2010, Principal Component Analysis, Wiley Interdisciplinary Reviews: Computational Statistics.
- Akdeniz, M. and Özdiñç, F.**, 2021, "Eğitimde Yapay Zeka Konusunda Türkiye Adresli Çalışmaların İncelenmesi", Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi,18(1),912-932., <https://doi.org/10.33711/yyuefd.938734>
- Aljedani, N., Alotaibi, R. and Taileb, M.**, 2020, "HMATC: Hierarchical multi-label Arabic text classification model using machine learning", Egyptian Informatics Journal.
- Aşliyan, R., Günel K. and Yakhno, T.**, 2007, Detecting Misspelled Words in Turkish Text Using Syllable n-gram Frequencies, In: Ghosh A., De R.K., Pal S.K. (eds) Pattern Recognition and Machine Intelligence. PReMI 2007, Lecture Notes in Computer Science, 4815. Springer, Berlin, Heidelberg.
- Beeferman, D., Berger, A. and Lafferty, J.**, 1999, "Statistical Models for Text Segmentation", Machine Learning 34, 177–210.
- Burkhardt, S. and Kramer, S.**, 2018, "Online multi-label dependency topic models for text classification", Mach Learn, 107, 859–886.
- Bozkurt, B. Ü.**, 2018, "Kavram, Kavramsallaştırma Yaklaşımları ve Kavram Öğretimi Modelleri: Kuramsal Bir Derleme ve Sözcük Öğretimi Açısından Bir Değerlendirme", Ankara Üniversitesi Dil Dergisi.
- Deepak, P., Visweswariah, K., Wiratunga, N. And Sani, S.**, 2012, "Two-part segmentation of text documents", Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12), Association for Computing Machinery, New York, NY, USA, 793–802.
- Duan, D., Qian, W. and Pan, S.**, 2012, "VISA: A visual sentiment analysis system", Proceedings 5th International Symposium Visa Information Communicate Interaction— VINCI'12, 22–28. ACM: Hangzhou.

KAYNAKLAR DİZİNİ (DEVAM)

- Ferrandin, M. and Cerri, R.,** 2023, Multi-label classification via closed frequent labelsand label taxonomies, *Soft Computing* (2023) 27:8627–8660, <https://doi.org/10.1007/s00500-023-08048-5>
- Geron, A.,** 2019, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly, 2nd Edition.
- Gör, İ.,** 2020, *Diferansiyel Denklemlerin Yapay Sinir Ağları İle Nümerik Çözümleri*, Doktora Tezi, Aydın Adnan Menderes Üniversitesi.
- Günel, K.,** 2006, *Intelligent Tutoring Systems for Education*, Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Fen bilimleri Enstitüsü, İzmir.
- Günel, K., Polat,R. and Kurt, M.,** 2016, Analyzing Learning Concepts in Intelligent Tutoring Systems, *The International Arab Journal of Information Technology*, 13(2).
- Heris, M. K. S.,** 2015, *Particle Swarm Optimization (PSO) in MATLAB*, Yarpiz Academic Source Codes and Tutorials.
- Hoon, G.K., Keong P.K. and Kong, T.E.,** 2006, “A Semantic Learning Approach for Mapping Unstructured Query to Web Resources”, *IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06)*,494-497, <https://doi.org/10.1109/WI.2006.24>.
- Hoon, G.K. and Wei, T.C.,** 2017, “Flexible facets generation for faceted search”, *First EAI International Conference on Computer Science Engineering EAI 1–3 Penang: Malaysia*.
- Jia, B. B. and Zhang M. L.,** 2023, Multi-dimensional multi-label classification: Towards encompassing heterogeneous label spaces and multi-label annotations, *Pattern Recognition*, 138 (2023), 109357.
- Jolliffe, I.,** 2005, *Principal Component Analysis*, *Encyclopedia of Statistics in Behavioral Science*, Citations: 648.

KAYNAKLAR DİZİNİ (DEVAM)

- Karaboga, D., Akay, B.,** 2009, A Survey: Algorithms Simulating Bee Swarm Intelligence, *Artificial Intelligence Review*, vol. 31 no., 1-4, 61-85.
- Kayahan, S., Günel, K. and Nuriyev U.,** 2022, Eğitim İçerikleri İçin Sezgisel Metin Bölütlemeye Dayalı Çoklu Etiketleme Stratejisi: MEB Sanat Tarihi Kitabı İçin Bir Durum Çalışması, *Bilişim Teknolojileri Dergisi*, Cilt: 15 Sayı: 2, 139 – 148.
- Kayahan, S., Günel, K. and Nuriyev U.,** 2023a, Multi-Labeling Strategy Based On A Heuristic Text Segmentation For Educational Content, *Journal of Modern Technology and Engineering* Vol.8, No.1, 2023, pp.39-53.
- Kayahan, S., Günel, K. and Nuriyev U.,** 2023b, A Clustering Method with Multi-labeling Using Particle Swarm Optimization for Texts in Turkish, II. International Scientific Conference, Ministry of Science and Education of Azerbaijan Republic Sumgayit State University. April 25-26, 2023.
- Kennedy, J. and Eberhart R.,** 1995, Particle swarm optimization, *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1942-1948 pp. vol.4, doi: 10.1109/ICNN.1995.488968.
- Kim, K.,** 2009, Face Recognition using Principle Component Analysis, Department of Computer Science University of Maryland, College Park.
- Kim, S., Lee, M. and Seok, J.,** 2022, Multi-label Text Classification of Economic Concepts from Economic News Articles using Natural Language Processing, Thirteenth International Conference on Ubiquitous and Future Networks (ICUF), DOI: 10.1109/ICUFN55119.2022.98.
- Kumar, V., Pujari, A.K., Padmanaphan, V., Sahu, S. K. and Kagita, V.R.,** 2018, “Multi-label classification using hierarchical embedding”, *Expert System with Application*, 91, January 2018, Pages 263-269.
- Kumova, S. and Karaoğlan, B.,** 2017, Stop Word Detection As A Binary Classification Problem, *Anadolu University Journal of Science and Technology*, 18(2): 346-359.

KAYNAKLAR DİZİNİ (DEVAM)

- Lee, J., Yu, I, Park, P. and Kim, D.W.**, 2019, “Memetic feature selection for multilabel text categorization using label frequency difference”, *Information Sciences*, 485, 263-280.
- Liu, J. Kang, J. Yu, and Z. Wang**, 2005, “A comparative study on unsupervised feature selection methods for text clustering”, In *Proc. 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 597–601.
- Luckin, R. and Cukurova, M.**, 2019, “Designing educational technologies in the age of AI: A learning sciences-driven approach”, *British Journal of Educational Technology*, 50 (6)2824-2838, 2019.<https://doi.org/10.1111/bjet.12861>
- Lukacs, G.**, 1951, *Estetik I*, Payel Yayinevi, Çeviren: Ahmet Cemal.
- MacQueen, J. B.**, 1967, Some methods for classification and analysis of multivariate observations, *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 281-297.
- Manning, C., Raghavan P. And Schütze H.**, 2009, *An Introduction to Information Retrieval*, Cambridge University Press.
- Merwe, D. W. and Engelbrecht, AP.**, 2003, *Data Clustering Using Particle Swarm Optimization*, IEEE 0-7803-7804-0 103
- Moyano, J. M., Gibaja, E. L., Cios K. J., and Ventura, S.**, 2018, “Review of ensembles of multi-label classifiers: Models, experimental study and prospects”, *Information Fusion*, 44, 33-45.
- Nabiyev, V. and Erümit, A. K.**, 2020, *Eğitimde Yapay Zeka: Kuramdan Uygulamaya*, Pegem Akademi Yayıncılık.
- Oh, H., Myaeng, S.H. and Jang, M.G.**, 2007, “Semantic passage segmentation based on sentence topics for question answering”, *Information Science (Ny)*, 177, 3696–3717.
- Ouyang, F. and Jiao, P.**, 2021, “Artificial intelligence in education: The three paradigms”, *Computers and Education: Artificial Intelligence*, 2, 100020. <https://doi.org/10.1016/j.caeai.2021.100020>.

KAYNAKLAR DİZİNİ (DEVAM)

- Pak, I. and Teh, L. P.**, 2017, “Text Segmentation Techniques: A Critical Review”, Springer International Publishing.
- Pearson, K.**, 1901, On Lines and Planes of Closest Fit to Systems of Points in Space, Philosophical Magazine. 2 (11): 559–572. doi:10.1080/14786440109462720.
- Pehlivanoglu, Y. V.**, 2017, Optimizasyon, Temel Kavramlar ve Yöntemler, https://ae.ieu.edu.tr/documents/optimizasyon_volkan.pdf Erişim tarihi: 15/12/2022
- Renz, A. and Hilbig, R.**, 2020, “Prerequisites for artificial intelligence in further education: identification of drivers, barriers, and business models of educational technology companies”, International Journal of Educational Technology in Higher Education, 17, 14 (2020). <https://doi.org/10.1186/s41239-020-00193-3>
- Reynolds, C. W.**, 1987, Flocks, Herds, and Schools: A Distributed Behavioral Model, Published in Computer Graphics, 21(4)
- Sarıman, G.**, 2011, Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi.
- Sinaga, K. P. and Yang, M. S.**, 2020, Unsupervised K-Means Clustering Algorithm, IEEE Access.
- Su, Z., Kogan, J. and Nicholas C.**, 2010, Text Mining: Application and Theory, M. W. Berry and J. Kogan (Eds), John Wiley & Sons, Ltd.
- Tarekegn, A.N., Giacobini, M. and Michalak, K.**, 2021, “A review of methods for imbalanced multi-label classification”, Pattern Recognition, 118.
- Türk Dil Kurumu**, <https://sozluk.gov.tr/> Erişim tarihi: 26/07/2023
- Wentao, C. Q. G. and Scheepers, C.**, 2017, “Effects Of Text Segmentation On Silent Reading Of Chinese Regulated Poems: Evidence From Eye Movements”, The Journal of Chinese Linguistics, 44, no.2 (June 2016): 265-286.

KAYNAKLAR DİZİNİ (DEVAM)

- Yang S. X. and Li, C. H.**, 2010, A Clustering Particle Swarm Optimizer for Locating and Tracking Multiple Optima in Dynamic Environments, Ieee Transactions On Evolutionary Computation.
- Yang, X. S.**, 2009, Firefly Algorithms Formultimodal Optimization, Proceedings of the Stochastic Algorithms: Foundations and Applications, Lecture Notes in Computing Sciences, vol. 5792, 178-178, Springer.
- Yang, Z. and Liu, G.**, 2019, “Hierarchical Sequence-to-Sequence Model for Multi-Label Text Classification”, IEEE Access, 7, 153012-153020.
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J. B., Yuan, J. and Li, Y.**, 2021, "A Review of Artificial Intelligence (AI) in Education from 2010 to 2020", Complexity, Article ID 8812542, 18 pages.
<https://doi.org/10.1155/2021/8812542>



TEŐEKKÜR

Doktora alıőmamın her aőamasında yardım ve öneriyle bana destek olan, engin bilgi ve tecrübeleriyle yol gösteren deęerli danıőman hocalarım Prof. Dr. Urfat NURİYEV ve Do. Dr. Korhan GÜNEL'e, tez sürecinde yanımda olup bilgi ve önerileriyle bana yeni bakıő açıları kazandıran Do. Dr. Refet POLAT'a ve hayatım boyunca kalbimde sevgi ve desteklerini hissettięim aileme teőekkürlerimi sunarım.

20 / 10 / 2023

İmzası

Selcan KAYAHAN

ÖZGEÇMİŞ

2009 yılında Balıkesir Üniversitesi Necatibey Eğitim Fakültesi Bilgisayar ve Öğretim Teknolojileri Eğitimi bölümünden mezun oldu. Aynı yıl Milli Eğitim Bakanlığı'na bağlı Bilişim Teknolojileri öğretmeni olarak göreve başladı. 2018 yılında Ege Üniversitesi Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü'nde yüksek lisans eğitimini tamamladı. Aynı yıl Ege Üniversitesi Fen Bilimleri Enstitüsü Matematik Anabilim Dalı Bilgisayar Bilimleri Bilim Dalı'nda doktora eğitimine başladı. 2017 yılından beri İzmir – Karşıyaka Aydoğan Yağcı Bilim ve Sanat Merkezi'nde Bilişim Teknolojileri öğretmeni olarak görev yapmaktadır.

YAYINLAR

- İpek, J., Kayahan, S., Karasu, M., Çukurbaşı, E. and Yeşil, E.,** 2013, Inspection of Techno-Pedagogical Educational Qualifications Of Mathematics Teacher Candidates, Procedia Social and Behavioral Sciences, Barcelona, Spain.
- Kayahan, S., İnceoğlu, M. M.,** 2013, Foreign Language Learning by Mobile Systems: A Review, 1st International ELT Conference, Gazi Üniversitesi Eğitim Fakültesi, 15 – 16 Kasım 2013, Ankara.
- Kayahan, S.,** 2014, Students Opinions on Teaching English Through Mobile Devices, 8. International Computers And Instructional Technologies Symposium, Trakya Üniversitesi Eğitim Fakültesi, 18–20 Eylül 2014, Edirne.
- Kayahan, S., Çırakoğlu, M.,** 2015, Bilişim Teknolojileri Ve Yazılım Dersinde Teknopedagogik İçerik Bilgisi İle Desteklenmiş Çoklu Zekâ Kuramı Uygulamalarının Erişmeye Ve Hatırlamaya Etkisi, 9. International Computers And Instructional Technologies Symposium, Afyon Kocatepe Üniversitesi Eğitim Fakültesi, 20–22 Mayıs 2015, Afyon.
- Kayahan, S., Özduran, K.,** 2016, İngilizce Dersinde Uygulanan EBA Market Mobil Yazılımlarına İlişkin Öğrenci Görüşleri, Akademik Bilişim Konferansları 2016, Adnan Menderes Üniversitesi Eğitim Fakültesi 3–5 Şubat 2016, Aydın.

- İnceođlu, M. M., Kayahan, S.,** 2016, «Dyner English For Success» Yazılımı İin Geliřtirilen Bařarı Testinin Geerlik ve Gvenirlik alıřması, 3.rd International New Trends in Education, Ege niversitesi, Baheřehir niversitesi & Georgia State University, 26–29 Nisan 2016. Seferihisar, İzmir.
- Kayahan, S.,** 2017, Dyner English for Success Eđitim Yazılımının Uygulanıřına İliřkin đretmen Grřleri, Yksekđretimde Eđitim Arařtırmaları Kongresi, Marmara niversitesi 19-20 Mayıs 2017, İstanbul.
- Kayahan, S., İnceođlu, M. M.,** 2018, Lise đrencileri ile Dyner Mobil Yazılımının Uygulanması: đretmen Grřleri, 12. International Computers And Instructional Technologies Symposium, Ege niversitesi Eđitim Fakltesi, 2–4 Mayıs 2018, zdere, İzmir.
- Kayahan, S., Nuriyev, U.,** 2018, Technology education in science and art centers (BILSEM) in Turkey, International Conference On Sustainable Development and Actual Problems Of Humanitarian Sciences, 14-15 Mayıs 2018. Azerbaijan University, Bak.
- Kayahan, S.,** 2018, Eđitim Uygulamalarında Yapay Zeka: Bir Derleme, 6Th International Instructional Technologies Symposium, Trakya niversitesi Eđitim Fakltesi, 12-14 Eylül 2018. Edirne.
- İnceođlu, M. M., Kayahan, S.,** 2018, Developing Academic Achievement Test For Dyner Mobile Education Software, EDULEARN18 Proceedings, Palme de Mallorca, Spain.
- Kayahan, S.,** 2018, Dyner Mobil İngilizce eđitim yazılımının, đrenenlerin akademik bařarılarına ve algılanan đrenmelerine etkilerinin incelenmesi, Yksek Lisans Tezi, Ege niversitesi Eđitim Fakltesi Bilgisayar ve đretim Teknolojileri Anabilim Dalı, İzmir.
- Kayahan, S., Bozok, G., Nuriyev, U.,** 2022, Evaluation Of Science and Art Centers (Bilsem) Summer School Robotics Workshops in Turkey, Information Systems and Technologies: Achievements and Perspectives, December 8-9 2022, Baku.
- Kayahan, S., Gnel, K. and Nuriyev U.,** 2022, Eđitim İerikleri İin Sezgisel Metin Bltlemeye Dayalı oklu Etiketleme Stratejisi: MEB Sanat Tarihi Kitabı İin Bir Durum alıřması, Biliřim Teknolojileri Dergisi Cilt 15, Sayı 2, 139-148.

Kayahan, S., Günel, K. and Nuriyev U., 2023, Multi-Labeling Strategy Based On A Heuristic Text Segmentation For Educational Content, Journal of Modern Technology and Engineering Vol. 8, No. 1, 2023, pp.39-53.

Kayahan, S., Günel, K. and Nuriyev U., 2023, A Clustering Method with Multi-labeling Using Particle Swarm Optimization for Texts in Turkish, II. International Scientific Conference, Ministry of Science and Education of Azerbaijan Republic Sumgayit State University. April 25-26, 2023.

